nature portfolio

Peer Review File

Indigenous American Genomes Reveal Unique Diversity and Evolution

Corresponding Author: Professor Tábita Hünemeier

Any redactions in this file are there to maintain patient confidentiality, the confidentiality of unpublished data, or to remove third-party material.

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Referee #1

(Remarks to the Author)
Overview:
This manuscript analyzes 128 high-coverage Indigenous American genomes from eight Latin American countries. The authors propose that (1) Indigenous American groups have high levels of novel genetic diversity, (2) genetic variation closely tracks geographic distribution, (3) three major dispersals shaped South American populations, and (4) there is evidence of Australasian and archaic (Neanderthal/Denisovan) introgression. While the data set is valuable, the primary conclusions are not convincingly supported by the current analyses and figures. Substantial reorganization, better figure clarity, more robust statistical testing, and clearer comparisons with previous research are necessary to strengthen the paper.

Major comments:
1. Relation to Previous Work
o The study does not cite Gusareva et al. ("From North Asia to South America…"), which makes similar observations on genetic diversity and selection signals in Indigenous populations. The authors should clarify which findings in the current paper go beyond or differ from Gusareva et al., especially regarding ancient gene flow and demographic history.
2. "Higher-Than-Expected Novel Diversity"
o The manuscript claims unexpectedly high novel genetic variation. However, the reported ~1.5 million novel variants in 128 individuals is lower than ~5 million in 180 African individuals or even the model's predicted ~3.7 million. Despite the authors' statement that "this estimate is in the same order of magnitude," the actual observations appear to contradict the assertion of higher diversity.
o The rarefaction plot in Figure 1 (mentioned in the text) shows Indigenous American samples having the lowest genetic variation among the populations compared. This contradicts the claim of higher-than-expected novel diversity and should be addressed directly.
3. Geography-Mirrored Genetic Variation
o The authors attempt to argue that genetic variation reflects geographic distance, akin to Novembre et al.'s "Genes mirror geography." However, Figures 2B and 2C are barely legible, with inconsistent labeling (e.g., "Guana-Chane" vs. "Guana-Chane_Arawak"). Similarly, Figure 3 is cluttered and labeled inconsistently.
o A key premise is based on UMAP, which has been criticized for its capacity to preserve geographic structure. The PCA in the supplement, which might offer a more standard approach, does not have population labels indicating the proposed geographic gradient.
o The AMOVA result reported ($p = 0.04$) is marginal and is not corrected for multiple comparisons, further weakening the claim.
4. Evidence for Three Main Dispersals
o The "third dispersal" into South America is based on weak evidence: a purported shared affinity between Indigenous South Americans and Ceramic Period Caribbean individuals. Residual structure from previous dispersals (or a batch effect) could explain this pattern.
o Figure 4, which underpins these claims, suffers from overlapping panels (A and B) and missing branches. Labels in the map portion (panel B) are too small and overlap with icons. Panel C's axes and bar plots are difficult to interpret. In Figure 5,

the “summary model” is unclear regarding what the “expected statistics” are, and the statistical test used to declare “no significant deviation” is not described.
o Overall, the interpretation of a novel dispersal would be more convincing if the authors performed formal tests (e.g., simulations) to show the data cannot be explained by simpler one- or two-wave models.
5. Australasian and Archaic Hominin Affinity
o The authors demonstrate some affinity between Indigenous Americans and both Australasian and archaic hominin DNA. However, the timing frameworks need clarity. For instance, sections discussing runs of homozygosity (ROH) are confusing: long ROH typically indicates recent inbreeding, whereas significant proportions of ROH suggest historical inbreeding. It is unclear how ROH overlaps are interpreted as evidence for gene flow events that occurred ~10,000 years ago.
o Maintaining a low level of “Ypykuéra” (Australasian-like) ancestry for 10,000 years might be entirely plausible under neutral evolution; simulations could clarify whether this persistence truly exceeds neutral expectations.
o The use of PBS to identify genomic regions under selection that also show Australasian affinity should be explained more carefully. PBS is designed to detect high differentiation from closely related populations, so its combination with distant affinity signals is non-trivial and needs more explicit rationale.
6. Ambiguity in Archaic Introgression
o In discussing archaic introgression, the manuscript suggests that “we were unable to determine whether introgression signals were associated with Indigenous American or European ancestral components.” This ambiguity leaves the main conclusions about adaptive introgression in question. Addressing this would require clearer partitioning of ancestry components in the analysis.

Minor comments:
1. Unclear Relevance of Arapium Language
o On page 8, the paper discusses the Arapium individual’s spoken language but does not integrate this into the broader genetic narrative. Explain how this linguistic note informs the study’s conclusions.
2. Data Accessibility
o The manuscript states that data will be available but provides no details. Specify the repository or timeline for data release.
3. Inconsistent Terminology
o The text consistently mentions “Indigenous Americans”, but the figures toggle between “Native Americans” and "Indigenous Americans". Use consistent terminology throughout.
4. Figure Readability
o Figures 2, 4, 5, and many Extended Data (6–13) have small fonts, overlapping elements, and unclear legends.
o Figure 1 labels are in lowercase, whereas the text uses uppercase.
o Figure 4: Panels A and B overlap; some branches are missing, and map icons are unreadable. Panel C’s axes lack clear labeling.
5. Statistical and Methodological Clarity
o Provide explicit test statistics, effect sizes, or p-values when describing “unexpected” signals. Currently, the reader cannot confirm whether findings are robust or marginal.
o Clarify if Arctic/North American populations were excluded fully or just from specific analyses.
6. Typos and Formatting
o Line 529: “bYpykuéra” should read “by Ypykuéra.”
o Line 1353: Add a space between the final paragraph of “Archaic introgression inference” and “Overrepresentation Analysis.”
o Figure 5B: The x-axis label “date” has no units. Indicate if it is in years.

Referee #2

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

Referee #3

(Remarks to the Author)
I have finished reviewing the work “Indigenous American Genomes Harbor Unique Genetic Diversity and a Complex Evolutionary Past” by Castro e Silva and Nunes et al.

The manuscript describes the generation and analysis of the largest compendium to date of whole genomes from Indigenous Americans. This is a remarkable effort led and developed mainly by American (in the sense of “from The Americas” and not “from the USA”) researchers. The study integrates the generated as well as previously available ancient and modern genomic data to gain numerous insights about the genetic variation and population history of the Indigenous peoples of the continent. They reveal the extent of new genetic variation observed in these genomes and its correlation with geography. They also show evidence and dating for three main dispersals into South America, signals of adaptive evolution, patterns of archaic introgression (adaptive and non-adaptive), and refinement of the genetic signal of Austronesian ancestry in some Indigenous Americans.

The amount of work and new insights is remarkable. This newly generated genomic data will advance many relevant studies on Indigenous populations of the Americas. I, however, find that the current version of the manuscript is very difficult to follow. The text, although generally well written, is extremely long, vague in some occasions and repetitive an intricate in others. A much-needed discussions on the ethical aspects, benefit sharing, and population engagement are lacking. The

current structure and lack of important detail prevented me from making a more technical review (although I do have few technical comments). I hope that my comments help the methodological part be more transparent in a future version and then these other technical aspects can be more meticulously reviewed.

General comments:
-Capitalize Indigenous throughout the text (sometimes it is and sometimes isn't).

-The generated data does allow for some level of Structural Variation (SV) characterization. It is fair if the authors decided not to avaluate it, but it should be at least acknowledged that a lot of new variation could be potentially found by looking at SVs too.

-The generated genomic data comes from different sequencing platforms. It is known that generating data from different platforms can generate biases (e.g. making individuals sequenced or genotyped on the same platform artificially closer to each other). How did the authors check this didn't influence their analysis?

-The main text and/or methods should give an overview of the ethical aspects of the sampling. This can be detailed in supplementary, but acknowledgment and reassurance that the work was culturally sensitive, ethical, and conducted following local regulations, should be made explicit early.
-Ideally, a repository with documented scripts/commands used to analyze the data should be provided. This is key for reproducibility.

-A general difficulty when reading the text is the mixed use of terms to refer to populations: sometimes by language group, sometimes by genetic cluster, sometimes by population name or broad region (eg. Aridoamerica) and does not always follow an intuitive criteria. Something similar happens when referring to temporalities (e.g. Late Pleistoscene, Middle/Late Holocene, pre Ceramic, Archaic, etc).

I mention some examples in the specfiic comments, but the text needs to be revised throughout with this in mind so the reader can understand better the geographic location of each.

-Why are the UpoA and UpopA2 not addressed at all in the text. Strangely, Mixe is the selected population for many f3 and D tests, when it is Mixe in fact the modern population with evidence of UpoA, but there is no acknowledgement of this or justification for selecting it as a representative of Mesomerica. I think how discussing how UpopA 1 and 2 play a role in their demographic reconstructions is important.

-Another feature that makes this manuscript difficult to follow at times is that the narrative follows an analyses-per-analyses flow. A more intuitive and engaging way to present the findings would be to discuss by topic or question to be answered, then the findings along with ALL the analyses that support them.

-Please don't refer to individuals as "Samples" (E.g. Figure 1) , this objectivizes the human beings that donated their samples. This wording is particularly problematic for Indigenous peoples in the Americas, who have been historically marginalized.

Specific comments:

Lines 203-205: More detail should be given (in Methods or Supplementary) about said probabilistic model. Enough detail to be able to reproduce the findings is needed.

224 – FROH is introduced to the unfamiliar reader for the first time here, but it is unclear what the F before ROH means.

Line 229- What is meant by genomic-geographic structure?

Lines 229-239. The first paragraph describes the four major South American clusters. Then the paragraph in 239 starts referring to Pima and Yaqui. Then Nahua-Xochitlán, Mixe … etc, without never mentioning that these are populations outside South America. Authors should not expect readers to know the exact geographic location of each population. Although the figures show location, having to go constantly from text to figure to makes sense of the findings is very tiresome and not engaging. Also, it is unclear to what analysis or figure the first sentence of the paragraph is referring to.

266. Why is the signal of geneflow between Meso and South America unexpected? It has been reported previously.

270 – The title of this section does not seem to describe its content. At least the "lasting impact of colonization" is not described.

288-The mentioning of the Arapium individual (with no geographic context) feels very sudden and lacking of context.

301-Define IBD here since it is the first time it's mentioned.

328-330 This statement needs a reference, is it a figure in the manuscript? Is it a previous paper? And what's based on?

358- The inclusion of ancient "individuals" is mentioned with very little detail. A general description is needed to understand

the reach of the analyses that involve them. What type of data (SNP capture I assume from methods)? From what temporal/spatial range?

374-388 Starting with “Our discussion focuses on..” is a weird way to start a paragraph that is not part of the discussion. Ideally you start paragraphs stating what you want to discover and what analyses / dataset you use to do so, then give a general idea of how your analyses support the findings. Then you close with a brief explanation of what the findings entail in the big picture of the paper. One would expect in this paragraph that all tree dispersals are described in general terms. Then details on each can be given in the following lines. Instead, here only two are mentioned. This contrast with the expectation makes the next paragraph (390-401) very difficult to follow.

441- “An exception is Spirit Cave” -> Authors need to be careful when referring to an individual and give the necessary spatiotemporal context so readers can interpret the statement. Spirit Cave is a place not an individual. This is something that happens in other places in the text, please look for them and change accordingly.

528- The phrasing is confusing. What “to others” mean here?

609 – This is the first mention in the text of PBS, so it should be named in full here.

675 – Put the correct name for SPrime and add the reference. Some mentions of Software lack references. Please check that all the software is credited properly.

719 – We -> they?

Figure 4 – The tree did not display correctly on my screen.

María C. Ávila Arcos

Referee #4

(Remarks to the Author)
The authors make a valuable and exciting contribution to understanding the genomic population history of South America and the entire Americas. The central element of the study is the presentation of 128 new, high-coverage Indigenous American genomes from 45 populations and 28 language families, spread across several Central and South American countries. These new data are an invaluable resource for a continent and its peoples that have been seriously understudied. When combined with previously published modern and ancient genomes, they thoroughly analyze the data to find new insights into how South America was populated, the processes shaping its population structure, and aspects of evolutionary adaptation. Their key findings include new knowledge about previously uncharacterized levels of genetic diversity in Indigenous South American populations, as well as support for models that suggest multiple independent dispersals into South America, followed by processes of differentiation, and the identification of candidate genes under selection linked to various traits.

The strength of this study lies in its thorough analysis of genomic data, utilizing state-of-the-art population genetic methods. Unique and novel insights—beyond the essential data provided—are demonstrated, for example, by the models that help better understand the “population-Y” or “Ypykuéra ancestry.” Here, the authors present the most compelling analysis to date of this ancestry component linked to Australasian populations. Additionally, the detailed analysis of genomic diversity provides critical empirical data to challenge the old “low genetic diversity paradigm” for Indigenous American populations. For these parts of the study, as well as the other valuable analyses, I see no potential issues; the methodology is solid, and the interpretations, to the best of my ability to assess, are supported by the data.

My critique of this manuscript is minor, and much of it may come down to style, personal preference, rather than essential issues. However, some points should be clarified, toned down, or adapted for the publication context. I want to start with a common critique of many “Novel Genome Collection” papers: it feels like the paper is overloaded with analyses, trying to cover every possible aspect, which sometimes leads to redundancies and sections that are harder to read. All analyses presented are valuable, but the authors missed an opportunity to contextualize their findings with the rich, interdisciplinary sources available for interpretation. There's a heavy focus on ancestry-related models, and the evolutionary discussions seem somewhat like a secondary addition. What is the significance of those findings? Don’t get me wrong—I think the observations are vital, but it would have been helpful if they had been discussed and placed within a broader context. The authors could have easily expanded this into 2-3 separate manuscripts. But again, this is a matter of style, and it certainly shouldn’t prevent this work from being published.

One aspect that should be addressed is the claim of “novelty” in some sections. While this study does a fantastic job in verifying several hypotheses previously discussed and by adding a new level of complexity to our understanding, it feels like prior work on South American population history is undervalued (including that of some of the co-authors). Examples include the discussion of how population structure was shaped in South America, as well as the EEMS analyses. This is indeed the most detailed study on that, but both aspects of isolation-by-distance and environmental and socio-cultural factors, which are considered significant in shaping population structure, are not new and have been discussed by other researchers before. Another example is (Line 267, and associated text parts) the observed increased migration rate between Mesoamerica and South America. Why is this unexpected? This has been discussed in the literature, and a recent ancient DNA study of Colombian ancestors has shown evidence of important periods of Mesoamerican gene flow into South America. I do not

mention these examples to devalue the importance of the author's study in any way. It just seems like claims of novelty should be somewhat toned down, or the work of others that addressed similar aspects should be made more visible.

The detection and description of the "third wave" of gene flow into South America is exciting; however, it sometimes seems that the way it is described could be misleading or confusing for the reader. For example, starting at Line 411, the authors state that "Almost all present-day indigenous South Americans… share a distinctive genetic affinity…." But later in the paragraph, they mention the exceptions, including the Andean Quechua speakers. A similar issue appears in Line 434, where it sounds like Quechua speakers are not Indigenous South Americans. It's important that the authors more carefully emphasize the geospatial aspects of the "third wave" expansion and influence. Regarding the Quechua speakers, in line 475, they mention that these groups received "…some gene flow from ancient Andean individuals." This could be misleading. Aren't the "ancient Andean individuals" the ancestors of the Quechua? Or do the authors mean that modern-day Quechua speakers received gene flow in addition to their ancestral DNA? Another key point to add to the "third-wave" discussion, especially concerning the admixture model in line 451, is a discussion of the limitations of using modern "Mixe" as a source. Using a current population introduces a weakness in the model because, due to the lack of ancient DNA data, this issue cannot be fully addressed. However, providing a brief overview of the population history of this source and explaining why the authors believe it is a valuable reference would be helpful.

Overall, this is an important and exciting study. The new data presented is invaluable, and the analytical results provided enhance our understanding of the evolution of Indigenous South American genomic diversity. Most of my critique here is minor and based on the approach I would take as an Anthropological Geneticist in describing the data and contextualizing the observations. None of this should prevent the publication of this excellent work. I recommend accepting this manuscript for publication with minor revisions.

Version 1:

Reviewer comments:

Referee #1

(Remarks to the Author)
Overview:
The revised study titled "Indigenous American Genomes Harbor Unique Genetic Diversity and a Complex Evolutionary Past" presents a genomic analysis of 128 high-coverage genomes from Indigenous American populations across eight Latin American countries. While the authors were incredibly loquacious in their 69 page response to reviewers, several comments were not sufficiently addressed.
1. We previously noted that the section "Indigenous American genomes reveal higher-than-expected novel diversity" was misleading. The authors argued that we misinterpreted what the expectation was, but I do not think this is the case. For example, the manuscript says "our results revealed a higher-than-expected genetic diversity compared to that of other continental populations". It later says "By analyzing 128 high-coverage genomes …, we identified 12,493,650 single nucleotide variants (SNVs), of which 11.4% were novel (1,426,511; Figure 1B)." In order to be consistent with the first quoted text, other continental populations should have a lower genetic diversity than this. Instead, the paragraph continues with "African populations exhibit approximately 5M (million) novel variants in 180 individuals17, while Oceanian populations show approximately 2M novel variants in 159 individuals." Clearly there are fewer novel variants among the Indigenous American genomes than Oceanian populations (1.4M vs 2M) with only a small change in sample size. The first quote is therefore inaccurate.
2. The model that the authors use to calculate an expected number of "~936,000 expected new variants" is based on an inaccurate model of Indigenous Americans. It does not appear to take into account the known population structure among all the groups that have existed since the Beringian bottleneck. The authors do make several claims about the bottlenecks and founder events that Indigenous Americans have faced, but do not at all comment on the rapid census expansions within the Americas. Regardless, if the authors developed an accurate demographic model of the Indigenous American populations, then they would recapitulate the observed number of genetic variants observed in the populations. Seeing a difference between observed data an the expectation of a model simply means the authors are using a poor model. Perhaps consider developing an improve demographic history using a tool like momi2 (https://www.tandfonline.com/doi/full/10.1080/01621459.2019.1635482).
3. Generally speaking, I do not think that genetic variation should be a bake off, whereby a population is valued because it contains more or less genetic variation than other populations. As proposed, the premise of this entire section seems somewhat troubling to me.
4. It does not seem like "Indigenous American genetic variation mirrors geographic distribution" is an appropriate section title for that section. While the PCA plot in Figure 2 does show the Pima at the top (representing "North"), it is not at all clear that that the vast majority of points follow the geographic pattern. For example, there are many groups sampled to the south of Surui, east of the Karitiana, and west and south of Amahuaca and Yaminahua. Yet these populations end up as the major labeled outliers. The neighbor joining tree in Figure 3 does not appear to convey geography. To argue that there is a geographic similarity in any of these plots, perhaps follow what was done in the Novembre et al (2008) paper that did this kind of analysis in Europeans? To this reviewer, it looks like the peopling of the Americas was a much more complicated process than the radiation in Europe, and potentially much more exciting to describe the complexity than to try to narrow it down to simple geography. Statements like "Analysis of molecular variance (AMOVA) suggested that genetic clusters corresponding to broad geographic regions accounted for a significant proportion of genetic variation ($p = 0.04$)" are not convincing. One of the newly added sentences captures my sentiment: "However, an isolation-by-distance model alone may

not fully explain the observed genetic variation across the region, as additional factors are likely contributing to the observed variation."
5. The authors seem to neglect the importance of the Gusareva paper. While in the rebuttal the authors claimed a difference in the number of populations sampled, they did not notice that several of the populations included in the present study were pooled into single groups in the Gusareva study. I think it would be helpful to acknowledge how the current study amplifies and clarifies some of the messages of the Gusareva et al paper and adds new insights, rather than attempt to belittle the paper with terms like "only".

Referee #2

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

Referee #3

(Remarks to the Author)
The authors have made an outstanding work to reply to the reviewers' comments. I think the improvement in clarity was much needed to make the manuscript more accessible. As mentioned by reviewer 4, these kind of papers tend to be loaded with data analyses and the essence of the message is often obscured by the methods and statistics. This paper is not an exception, but I believe it's a reflection of the field and should not demerit the valuable work and effort put in this particular manuscript.

I am satisfied with the response to most of my comments. I have very minor things that I think still require some attention.

1.- Both reviewer 1 and I raised concerns about the model to test whether the novel variation identified is a lot or "expected". In my opinion, the generation of this genomic data is an extremely valuable contribution on its own right, and an analysis to justify it is not needed (at least not in the main text). I think that stating that a lot of new variation is found is good enough. The space used to contrast against a given model (that could be in principle parametrized in many ways) is needless and distracts from the many other analyses and inferences that come later in the text. This is just a suggestion to be gentle on the readers and not saturate them beyond the minimum necessary.

In any case, if the authors decide to keep this part, along with the new additional information produced in response to Reviewer 1, then more discussion should be given as to why this is relevant. Is it because the demographic model is wrong? Or why?

2.- Reviewer 4 and I agree that the Meso -> South geneflow is not unexpected. I appreciate the added text to contextualize, but I think removing the word "unexpected" is still needed.

3.-The addition of a closing sentence to the section "Uneven genetic diversity distribution and the lasting impact of colonization" helps in giving more sense to the title. However, the text throughout the section does not seem to assert a cause-consequence (i.e. "impact") relationship. I think I missed that the authors were trying to, in fact, state the "lasting impact of colonization" because this was not explicitly or actively described as such.

Drawing from the sentences highlighted in the response by the authors:

"These groups also exhibited an excess of long ROH segments (Extended Data Figure 5), indicating recent inbreeding likely caused by population collapse, fragmentation, and post-colonization isolation" -> What caused this collapse?

"Additionally, identity by descent (IBD) analyses revealed that segments >8 cM were predominantly shared within populations (Extended Data Figure 6; Supplementary Table 8), suggesting limited gene flow since the colonial period. Therefore, these small and isolated populations likely experienced increased inbreeding during this period, leading to the high prevalence of long ROHs (Extended Data Figure 5)." -> What caused this limited gene flow?

"Effective population size (Ne) histories inferred from IBD analysis (methods), grouping individuals by genetic cluster, revealed a widespread post-European contact bottleneck across all Indigenous American populations, followed by signs of recent genetic diversity recovery in Chaco and western South America (Figure 3D)." -> What caused this "bottleneck"?

"In contrast, Aridoamerica exhibits persistently low Ne values and experienced the most pronounced post-European contact bottleneck (Figure 3D), consistent with the high amount of long IBD segments shared within this cluster (Extended Data Figure 6)." -> same as above.

I am certain that authors are aware and sensible to the lasting impact of colonial dominance on Indigenous Americas, the passages above, however, don't convey it. Referring to the Indigenous population collapse as a passive "demographic bottleneck" sanitizes history and obscures the known causes: epidemics, massacres, and displacement directly perpetrated by colonizers. The text as it stands reports findings accurately and soundly, but my very personal opinion is that historical accountability is important. This is just a personal recommendation and authors can decide what is best.

Referee #4

(Remarks to the Author)
The authors present a revised version of their manuscript, featuring 128 new, high-coverage Indigenous American genomes from 45 populations and 28 language families, distributed across several Central and South American countries. These new data are an invaluable resource for a continent and its peoples that have been seriously understudied, and the presented results represent a significant advancement in our understanding of South American population history. In my previous review of the study, I voiced some minor concerns regarding the contextualization of the authors' analyses with previously published studies and some minor issues with clarity. The revised manuscript effectively addressed all these critiques, and, along with the revisions based on other reviewers' comments, this manuscript has undergone substantial improvement. The authors now present their excellent analytical work in a comprehensive and thoughtful manner. I am very excited about this important work and recommend the manuscript for publication without further revisions.

Version 2:

Reviewer comments:

Referee #1

(Remarks to the Author)
Overall:
I am honored to have been a part of this review process. I think the revisions suggested by all reviewers have substantially strengthened manuscript, with all claims in the paper now being strongly supported with direct analyses, or toned down to reflect observations in the data. The work was conducted in collaboration with Indigenous communities, and data were shared back to the communities. This is an underappreciated, yet essential step for representative genomics projects that all studies should emulate. I have a few remaining minor comments that in my view should not hold up the manuscript with further review.

Section "Global genomic databases miss extensive Indigenous American diversity"
• I would recommend removing the sentence "Although we did not assess structural variation here, these genomes are likely to reveal an additional layer of previously uncharacterized structural diversity, including copy number variation and insertions and deletions." from the section. It is too speculative for a results section, and could be moved to the conclusions if the authors want to keep it.
• To me, the new text of this section jumps back and forth between two topics. I would recommend moving the sentence "By analyzing 128 high-coverage genomes against a dataset of more than 270,000 individuals from various genomic databases (1KGP21, HGDP13, SGDP12, GNOMAD22, and dbSNP23), we identified 12,493,650 single nucleotide variants (SNVs), of which 11.4% were novel (1,426,511; Figure 1B). Among these new variants, 4.7% were polymorphic (allele frequency (AF) > 0.01), and 0.02% were common (AF > 0.05; Figure 1C-E)." down to the second sentence of the second paragraph where the authors focus on novel variation, so that the first paragraph just talks about total variation and the second paragraph just talks about novel variation.

Section "Demographic history and geography leave distinct signatures in genetic variation"
• The conclusion that Suruí and Karitiana may be poor representatives for South American Indigenous ancestry in the new text is important and clearly laid out.

Referee #2

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

Referee #3

(Remarks to the Author)
I have no further comments and appreciate the efffort by authors to address the feedback.

RESPONSE TO REVIEWERS

REVIEWER 1

Dear Reviewer 1,

Thank you for your valuable comments. We believe that your suggestions greatly enhanced the quality of our work. We value your time, detailed perspective, and effort to help us improve our manuscript. Please find below your comments (marked in bold) and our replies following each one of them. The corresponding changes can be found in the manuscript attached with those marked in red.

Sincerely,

Tábita Hünemeier and colleagues

Comment 1. The study does not cite Gusareva et al. ("From North Asia to South America…"), which makes similar observations on genetic diversity and selection signals in Indigenous populations. The authors should clarify which findings in the current paper go beyond or differ from Gusareva et al., especially regarding ancient gene flow and demographic history.

Authors' reply: We would like to begin by thanking the reviewer for their careful evaluation and constructive comments on our manuscript. We have thoroughly considered and incorporated the reviewers' feedback, which we believe has significantly enhanced the quality and clarity of our work.

Regarding the recommendation to cite the study by Gusareva et al., we appreciate the reviewer bringing this to our attention. However, we would like to note that their paper was published after our manuscript was submitted, and we were not aware of its release at the time. Gusareva et al. work bears very few similarities to ours, as it has different focuses, includes only a limited number of populations from America (13 Indigenous-related groups plus 3 admixed American groups), and lacks any linguistic or anthropological context. We acknowledge the contribution of Gusareva et al. and agree that their study should be cited in relation to our findings, despite the limited opportunities for direct comparison. In the new version of our manuscript, we have included references to the comparable results and analyses in the text.

<u>Regarding the inferred split times between genetic clusters:</u>

Within the continent, genetic differentiation among Indigenous American clusters intensified during the Middle Holocene, becoming particularly pronounced over the Late Holocene (Figure 4C). This period coincided with the emergence of distinct ancestral components and the establishment of regionally structured genetic profiles (Figure 4C). Supporting this, the cross-coalescence rate between Indigenous South American clusters begins to rise sometime between 5,000 and 3,000 years ago for most cluster pairs (Extended Data Figure 7). These results contrast with an independent study that estimated the divergence of major Indigenous South American groups (Amazonia, Andes, Chaco, and Patagonia/Southern Cone) to have occurred between 13,900 and 10,000 years ago[37]. Our inferences align more closely with the known timeline of dispersals into South America, particularly the second major dispersal, whose individuals, believed to have contributed most substantially to present-day populations, only began arriving around 9,000 years ago[5,6], after the divergence times proposed by the referred study[37]. This discrepancy may reflect our broader and more representative sampling of South American Indigenous populations, in terms of both number and ethnolinguistic and geographic diversity.

Furthermore, our inferred timeframe aligns with the diversification of several language families[38–40] and provides additional support for the increasing population structure during this period. The Middle Holocene brought a major climatic shift, with reduced summer insolation in the Southern Hemisphere and increased insolation in the Northern Hemisphere, which weakened the South American monsoon and triggered climatic instability[41]. Along with the extinction of megafauna[42], these climatic changes disrupted ecosystems and reduced resource availability. Archaeological evidence indicates widespread demographic decline and retreat into ecological refugia, likely contributing to the genetic isolation and differentiation observed in our data[43]. Population recovery followed, likely boosted by early plant cultivation and the rise of regionally structured societies[43].

Regarding the demographic inference from coalescence rates:

Finally, we inferred the N_e history of genetic clusters based on within-population coalescence rates (see Methods) and found that over the past 10,000 years the effective population size of Indigenous Americans declined by approximately 40–90%, corroborating previous findings[37]. This analysis also revealed that historically, Indigenous North American populations exhibited slightly higher N_e than their South American counterparts (Figure 3E). However, in the last 1,000 years, this trend has reversed, with South American lowland populations (Western and Eastern South American clusters) displaying marginally higher N_e. This pattern aligns with the IBD-based N_e inferences, which indicate a recent population expansion in the region (Figure 3D–E). Notably, IBD-based methods are sensitive to recent demographic events (within the last 50–100 generations), but lack resolution for deeper timescales due to the unreliable detection of short segments. In contrast, coalescent-based approaches are more powerful at capturing ancient demographic events but lose resolution in the recent

past, where fewer lineages remain and thus fewer coalescent events provide information about population history. Therefore, integrating both improves the inferences across temporal scales. Taken together, these findings illustrate the complex demographic history of Indigenous Americans, from their divergence from Asian ancestors to their genetic differentiation across the American continent. They also shed light on the lasting impacts of the colonization process on the levels and distribution of genetic diversity in present-day populations.

Regarding the inference of positive selection:

We acknowledge that Gusareva et al. employed a methodological framework broadly similar to ours (iHS, XP-EHH, and PBS). However, there are substantial differences in the focal populations, the hypotheses tested, and consequently in the results obtained.

Gusareva et al., focused on detecting selection signals associated with specific ecological contexts (e.g., cold Siberian environments and high-altitude Andean settings). To this end, they analysed population groups defined through hierarchical clustering of individuals based on Admixture results, using a European population (GRB) as an outgroup in the PBS analysis. In contrast, our study aimed to identify selection signals shared across Indigenous populations throughout the Americas. To address this question, we used Siberian populations as the sister group and East Asian populations as the outgroup, which allowed us to detect selection signatures that occurred after the divergence of these groups within the American continent.

As a result, Gusareva et al. report for the Americas only the well-known signal in *EPAS1* associated with hypoxia adaptation in Andean populations—a finding already described in previous studies (e.g., doi.org/10.1038/s41598-017-13382-4). In our work, by focusing on shared selection across the continent, we identified both previously reported and new candidate genes related to a broader range of traits under selection, including immune response, cardiometabolic traits, fertility, and anthropometric features.

Therefore, while the methodological tools may overlap, the scope, study design, and interpretive framework differ substantially. With all due respect to the reviewer, we believe that a direct comparison of results may not fully reflect the distinct goals and contributions of each study, and we have therefore decided not to reference Gusareva et al. in the section discussing natural selection.

Comment 2. The manuscript claims unexpectedly high novel genetic variation. However, the reported ~1.5 million novel variants in 128 individuals is lower than ~5 million in 180 African individuals or even the model's predicted ~3.7 million. Despite the authors' statement that "this estimate is in the same order of magnitude," the actual observations appear to contradict the assertion of higher diversity.

Authors' reply: We appreciate the reviewer's concern. We realized that there was a misinterpretation of our text, since the inference of 3.7 million variants refers to an estimate without considering any demographic effects, as stated in the manuscript:

> This estimate arises from a probabilistic model that considers the chance of emergence of new variants throughout generations and does not account for population expansion following the peopling of the continent or the severe bottleneck caused by European colonization, which reduced the effective population size by at least 80%.

Redacted

To clarify this issue and make the text more coherent, we re-ran the analyses taking into account possible demographic changes over time. Our intention was to emphasize that the level of genetic diversity observed in these newly sequenced Indigenous American genomes is higher than what would be intuitively anticipated based on expectations from their demographic history (i.e., the relatively small Ne and the series of founder effects and bottlenecks associated with the initial peopling of the continent) as well as on previous studies, which primarily focused on a limited number of populations such as the Karitiana and Suruí (Bergström et al. 2020; Mallick et al. 2016). As we discuss in the manuscript, these specific groups (Suruí and Karitiana) present high levels of runs of homozygosity as a consequence of recent inbreeding, and are not representative of the broader genetic diversity found among Indigenous Americans, not even among South American lowlanders.

To provide evidence that the number of newly identified variants in Indigenous Americans ex ceeds expectations based on their demographic history, we have refined our estimates of the e xpected number of new variants, as presented in the section *"Indigenous American genomes r eveal higher-than-expected novel diversity."* *Redacted*

This finding likely reflects the persistent underrepresentation of Indigenous American populations in genomic research, resulting in much of their genetic diversity remaining uncharacterized. Previous studies have shown a positive correlation between the number of newly identified variants in Brazilian individuals and the proportion of their Indigenous American ancestry (Nunes et al. 2025).

Taking these points into account, we have revised and updated this section to clarify our statement that the number of novel variants observed exceeds expectations and to present the results of the updated analysis.

Redacted

Comment 3. The rarefaction plot in Figure 1 (mentioned in the text) shows Indigenous American samples having the lowest genetic variation among the populations compared. This contradicts the claim of higher-than-expected novel diversity and should be addressed directly.

Authors' reply: We appreciate the reviewer's observation and acknowledge that our original phrasing can lead to misinterpretation. As clarified in our response to the previous comment, our intent was to highlight that the genetic diversity identified in this newly analyzed set of Indigenous American genomes is greater than expected based on their demographic history and on earlier studies. A population may display the lowest levels of diversity among present-day groups and yet still harbor more diversity than expected under the models of the peopling of the Americas and the out-of-Africa expansion. This underscores the importance of incorporating empirical data from underrepresented populations to better understand the genetic evolution of human groups. In this sense, we did not mean to suggest that this diversity surpasses that of other continental groups; indeed, we are aware that Native American populations are genetically less diverse than other continental populations. However,

considering the expectations of both the out-of-Africa model and the bottlenecks associated with the Beringian dispersal, the levels of diversity we observe are higher than expected, as shown in the new analysis and explained in our response to the reviewer's comment 2. We have revised the manuscript to make this distinction clearer, as detailed in our response to Comment 2.

Comment 4. The authors attempt to argue that genetic variation reflects geographic distance, akin to Novembre et al.'s "Genes mirror geography." However, Figures 2B and 2C are barely legible, with inconsistent labeling (e.g., "Guana-Chane" vs. "Guana-Chane_Arawak"). Similarly, Figure 3 is cluttered and labeled inconsistently.

Authors' reply: We thank the reviewer for their comment and acknowledge that the relationship between geographic distribution and genetic similarity in humans has been well established before.

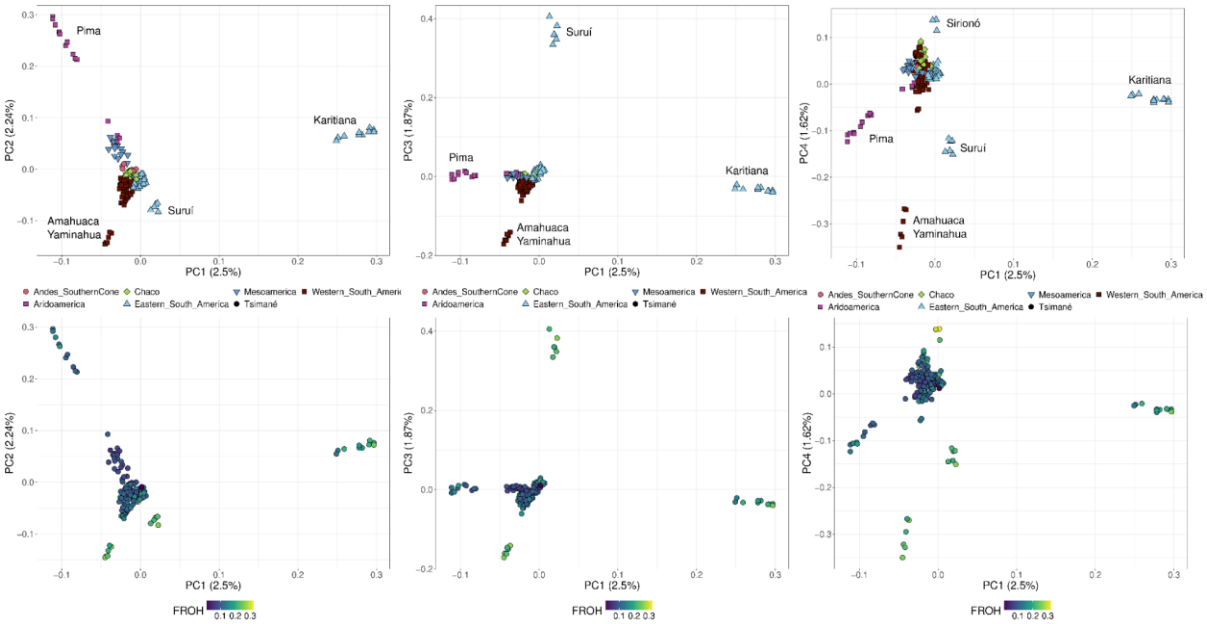
We apologize for the low resolution of the figures in the final PDF. This issue arose because the figures embedded in the manuscript file were automatically compressed by the text editor, resulting in reduced resolution. The original figures were prepared at high resolution and do not exhibit these artifacts. To address this, we have now provided a supplementary PDF containing high-resolution versions of all figures for the reviewers' reference.

Regarding the figure labeling, the noted inconsistency arises from the fact that in Figure 2B we used only the name of each ethnic group to identify the populations, whereas in Figure 2C we included both the ethnic group name and the corresponding language family in the format: [ethnic group]_[language family]. To ensure consistency throughout the figures, we have now standardized the labeling by using only the ethnic group names to identify all populations.

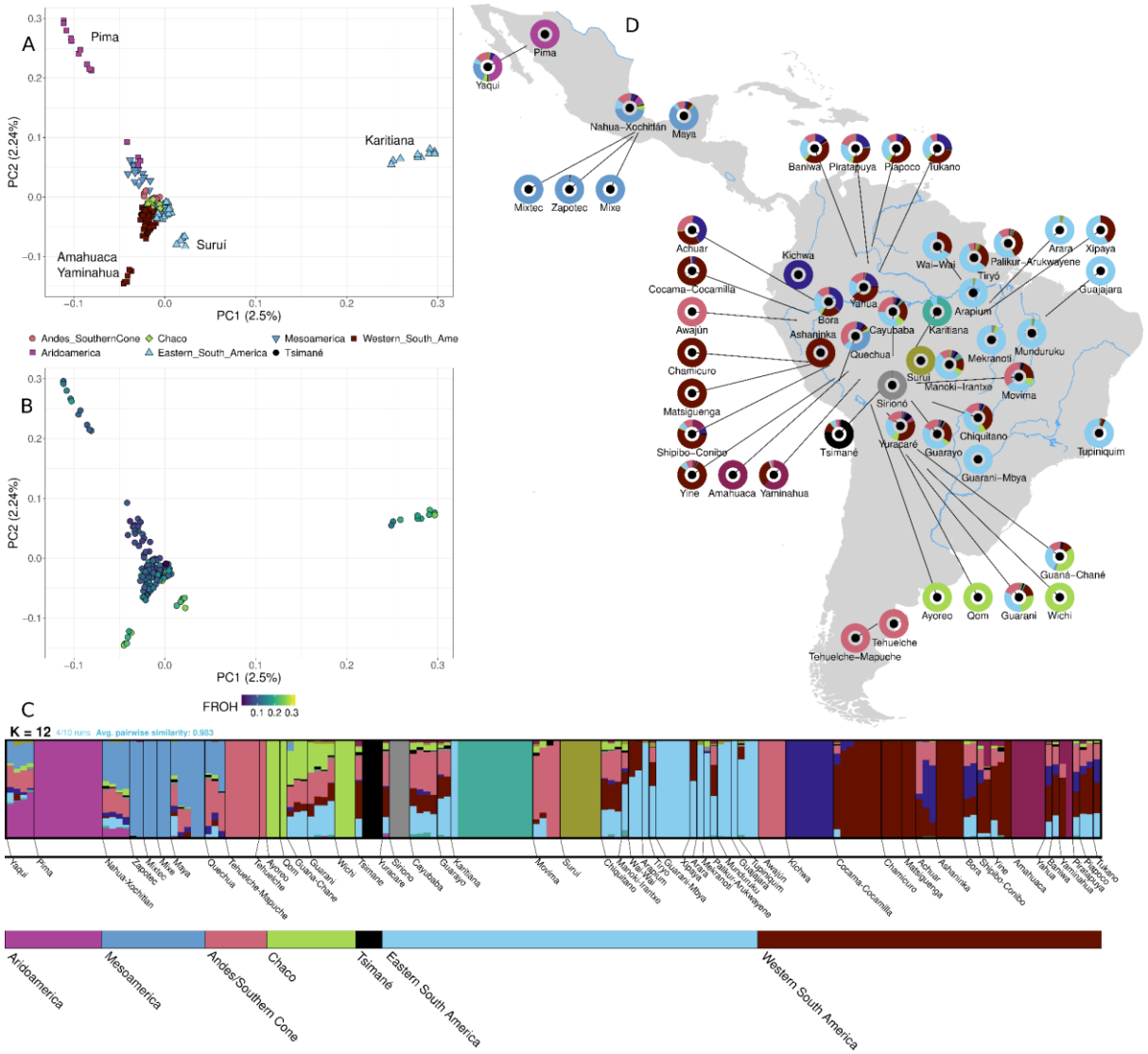
Comment 5. A key premise is based on UMAP, which has been criticized for its capacity to preserve geographic structure. The PCA in the supplement, which might offer a more standard approach, does not have population labels indicating the proposed geographic gradient.

Authors' reply: We thank the reviewer once again for their thorough evaluation of our manuscript's methodology and fully understand the concern raised. We wish to clarify, however, that the UMAP analysis was used solely as an initial exploratory tool and does not underpin any of our main conclusions. UMAP is primarily intended as a visualization technique, not as a method for statistical inference. We originally included the UMAP analysis to provide an overview of the overall genetic variation (summarizing all principal components into two dimensions) and to facilitate visualization of broad patterns. Specifically, we highlighted the moderate separation between North and South American genetic variation, and, more importantly, the observation that populations appearing as outliers along the principal components often exhibit the highest levels of runs of homozygosity (Figure 2A–B).

Importantly, the first principal components from the PCA reveal a pattern consistent with that shown in the UMAP (figure below). The most divergent groups are largely the same; however, in the PCA (specifically along PC2) the Pima appear as an outlier. Similar to the UMAP results, PC1 and PC3 primarily distinguish the Karitiana and Suruí as outliers relative to the genetic variation of other Indigenous American groups (upper panels of the figure). Additionally, the same trend observed in the UMAP analysis is evident in the lower panels, where individuals with higher FROH values exhibit greater differentiation.



Based on these observations, and to simplify both the analytical framework and interpretation, we replaced the UMAP results previously shown in Figure 2A–B with PC1 and PC2 from the PCA (see the updated figure below).



Additionally, it appears the reviewer might have interpreted our findings as relying on the UMAP analysis to support the conclusion of genetic differentiation between North and South America, as well as among different regions within South America (i.e. Southern Cone, Chaco, Eastern and Western South America). Specifically, the suggestion seems to be that we inferred a relationship between genetic structure and geographic distribution in Indigenous Americans based mainly on UMAP. In reality, this conclusion is primarily supported by the analysis of pairwise genetic distances calculated using the outgroup f3 statistic, which provides an unbiased measure of genetic similarity unaffected by population-specific drift (presented in Figure 3A). These distances were further examined using standard techniques widely employed in human population genetics, namely Neighbor-Joining trees and Multidimensional Scaling. These results are also corroborated by the ADMIXTURE analysis (Figure 2C), in which the genetic structure also mirrors the geographic distribution of populations. Finally, we also show that genetic and geographic distances between pairs of populations are significantly correlated (Supplementary Figure 3).

The results from these analyses, shown in Figure 3, form the basis for defining the genetic clusters referenced throughout the manuscript, including those depicted by color in Figure 2. We recognize that the sequence in which these analyses were introduced may have unintentionally caused confusion regarding which results support our conclusions and, for instance, how the genetic clusters were defined.

To address this, we have revised the legend of Figure 2 to prevent misinterpretation regarding how the genetic clustering was inferred and how cluster affiliation was defined.

> **Figure 2 - Genetic structure of present-day Indigenous American populations.** (A) The first two principal components (PCs) inferred from LD-pruned masked genome-wide data. Individuals are color-coded by genetic cluster, as defined in the analysis presented in Figure 3A. (B) PCA visualization of the same data, color-scaled by individual ROH-based inbreeding coefficients (F_{ROH}). (C) Ancestry proportions inferred via unsupervised ADMIXTURE analysis (the highest K for which a consensus was reached), ordered by genetic cluster (as defined in Figure 3A) and plotted with PONG. (D) Geographic distribution of population-averaged ancestry proportions at K = 12 overlaid on a partial map of America. Both analyses were restricted to the maximum set of individuals unrelated at or above the first-degree level.

We believe that presenting the exploratory data analysis and basic population structure inference methods first (namely PCA and ADMIXTURE; Figure 2) followed by the more detailed genetic clustering based on outgroup f3 distance metrics (Figure 3), offers a more coherent structure. This sequence aligns with standard practices and the typical organization found in most studies within the field. Nevertheless, we have revised and reorganized the section titled *"Indigenous American genetic variation mirrors geographic distribution"* to clarify this rationale and improve the interpretability of our findings. Please note that we have also integrated additional suggestions from the reviewers into the revised text of this section.

> We characterized present-day genetic structure and variation among Indigenous American populations using Principal Component Analysis (PCA) and ADMIXTURE (see Methods) on a masked dataset of 160 unrelated genomes. PCA of global reference populations revealed no evidence of batch effects, as individuals were consistently clustered by geographic origin rather than by dataset source (Supplementary Note 1). Our analyses revealed genetic differentiation between North and South American Indigenous groups, as well as substantial substructure within each region (Figures 2 and 3). Several populations, including the Karitiana, Suruí, Amahuaca, and Yaminahua, exhibited highly distinct genetic profiles, driven in part by increased levels of inbreeding, as evidenced by the higher proportion of the genome falling within runs of homozygosity (F_{ROH}; Figure 2; Extended Data Figure 3; Supplementary Table 4). These findings highlight the limitations of using the Suruí and Karitiana as representative proxies for lowland South American genetic diversity, given their pronounced genetic differentiation, even from geographically proximate populations.

This relationship between genetic similarity and geographic distribution becomes clearer through outgroup f3 statistics, which provide a measure of genetic similarity that is unbiased by population-specific genetic drift. A neighbor-joining (NJ) tree and multidimensional scaling (MDS) applied to 1–outgroup f_3 and 1/outgroup f_3 (Methods; Figure 3; Extended Data Figure 8A–B; Supplementary Table 5) revealed four major genetic clusters within South America that align closely with geographic regions. Accordingly, we named them based on their respective locations: (i) Southern Cone, (ii) Eastern, (iii) Western (mostly lowlands) South America, and (iv) Chaco. The Southern Cone populations (Mapuche-Tehuelche) were the most differentiated in South America, yet the Andean Quechua cluster was positioned within the same clade, suggesting a long-distance genetic connection. The Chaco populations form the sister branch, followed by Tsimané, which remains unclustered, possibly because of their cultural and linguistic isolation[25,26]. The final division separates Eastern and Western South American lowland populations.

Comment 6. The AMOVA result reported (p = 0.04) is marginal and is not corrected for multiple comparisons, further weakening the claim.

Authors' reply: We thank the reviewer for raising this important point. In line with the original formulation of AMOVA (Excoffier, Smouse & Quattro 1992; Excoffier & Lischer 2010), we report the permutation p-values without multiple-testing correction. AMOVA is conceived as a single hierarchical model partitioning genetic variance, and the significance values are intended to evaluate how much of the variance is explained by a given a priori grouping. This approach has been consistently used in population-genetics studies, where alternative grouping hypotheses (e.g., geography vs. language) are tested independently and compared by the proportion of variance explained rather than by applying post hoc multiplicity corrections. Moreover, in our study AMOVA is used as an exploratory analysis, whose outcome is corroborated and reinforced by a broader set of sequential tests presented in the main text. As outlined in the manuscript, multiple lines of evidence (including the ADMIXTURE analysis, the Neighbor-Joining tree and the MDS based on outgroup f3 distances, and the significant correlation between genetic and geographic distances) indicate a relationship between genetic structure and geography. Nonetheless, an isolation-by-distance model alone does not fully explain the spatial patterns of genetic differentiation. This is supported by the EEMS analysis (Extended Data Figure 4), which identifies regions with either elevated or reduced genetic differentiation compared to what would be expected under a simple isolation-by-distance model. Consequently, the lack of high statistical significance in the AMOVA is consistent with the notion that grouping populations solely by geographically structured genetic clusters does not fully capture the complexity of genetic variation across the region.

Analysis of molecular variance (AMOVA) suggested that genetic clusters corresponding to broad geographic regions accounted for a significant proportion of genetic variation ($p = 0.04$), whereas grouping by ethnolinguistic units (language families) had not such effect ($p = 0.57$; Supplementary Table 6). Additionally, genetic

distances, measured as $1-f_3$(Mbuti; X, Y), are significantly correlated with great-circle geographic distances between all pairs of present-day Indigenous American populations (Spearman's r = 0.4401, p ≅ 0 for all Indigenous Americans; r = 0.148, p ≅ 0 for South America alone; Supplementary Table 7 and Supplementary Note 4). However, an isolation-by-distance model alone may not fully explain the observed genetic variation across the region, as additional factors are likely contributing to the observed variation. In line with this, the estimated effective migration surface (EEMS) models indicated regions with genetic differentiation greater than expected under an isolation-by-distance model, notably southeastern Peru, northern Bolivia, and southwestern Amazonia (Extended Data Figure 4). This pattern closely resembles findings from a previous study involving a more limited set of populations from western South America[29]. Lower inferred migration rates were observed in Aridoamerica, the northern Peruvian and Ecuadorian Amazon, the Chaco region, and eastern South America, especially in areas north of the Amazon River. These patterns suggest that demographic history and geographic and cultural factors (including topography, climate, river systems, vegetation density, and sociocultural boundaries) beyond geographic distance have also shaped genetic differentiation. Additionally, an unexpected signal of increased migration rate between Mesoamerica and South America indicates greater genetic similarity than geographic distances alone would predict. This observation aligns with previous findings indicating gene flow from North and Central American populations into South American groups[5,33].

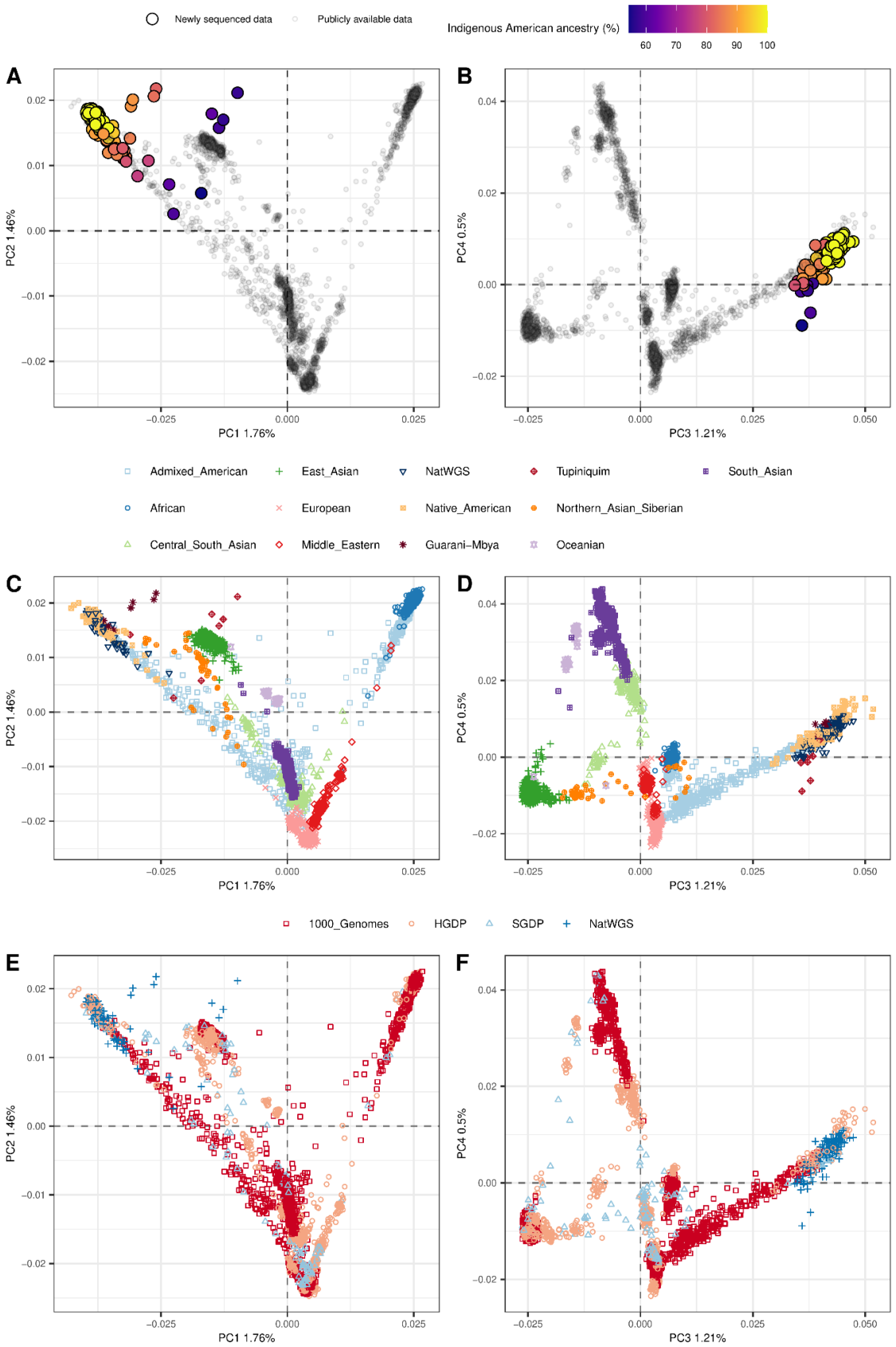
Comment 7. The "third dispersal" into South America is based on weak evidence: a purported shared affinity between Indigenous South Americans and Ceramic Period Caribbean individuals. Residual structure from previous dispersals (or a batch effect) could explain this pattern.

Authors' reply: We thanks to the reviewer for raising this important question, which addresses one of the key findings of our study. Our analyses consistently demonstrate that the genetic affinity between most present-day Indigenous South American populations and the Ceramic Period ancient Caribbean individuals is robust across datasets, marker sets, and methods, and therefore unlikely to be explained by batch effects or residual structure from earlier dispersals. This result is replicated in PCA, admixture graphs, qpAdm models, and qpWave (new formal analysis; see response to comment 9), all of which converge on the need for a distinct ancestry stream linking Mesoamerican-related sources with present-day South Americans and Ceramic Period Caribbean populations. To lend further support to this conclusion, we also systematically investigated potential technical and demographic biases that could produce such a signal, as detailed below.

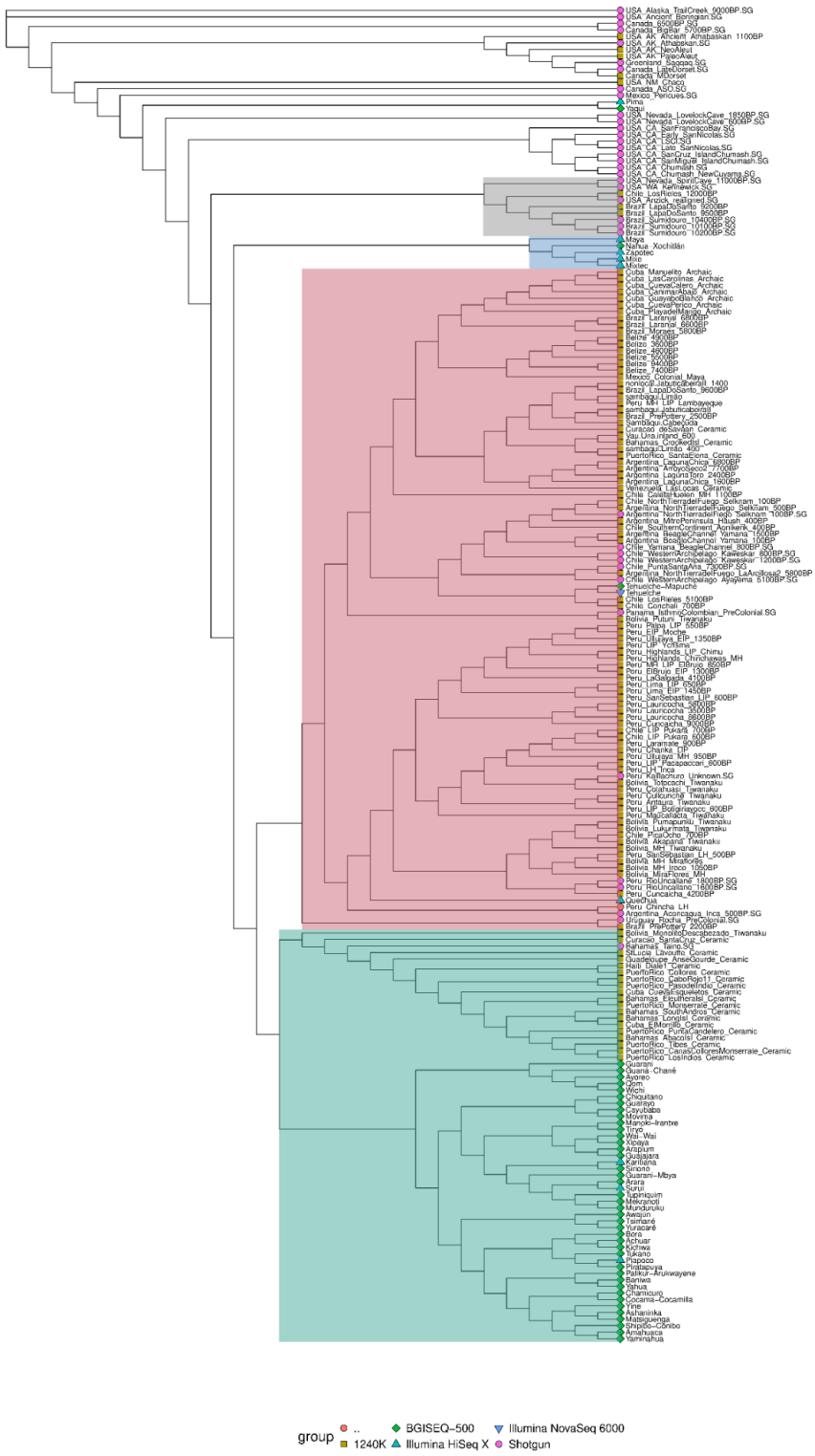
We performed a PCA including all present-day individuals (presented in Supplementary Figure 2 in Supplementary Note 1), comprising both the 128 newly sequenced genomes and individuals from public datasets (1000 Genomes Project, Human Genome Diversity Project,

and Simons Genome Diversity Project). As noted in the main text, “PCA of global reference populations revealed no evidence of batch effects, as individuals were consistently clustered by geographic origin rather than by dataset source (e.g., HGDP, SGDP, or 1KGP), indicating robust integration across datasets (Supplementary Note 1).”

In this analysis, the newly sequenced individuals (dark blue) cluster closely with publicly available Indigenous American genomes (light orange), supporting the absence of batch effects among Indigenous American individuals (Supplementary Figure 2). The few outliers observed correspond to individuals with higher proportions of non-Indigenous American ancestry (see panels A and B in the figure below), such as those from the Tupiniquim and Guarani-Mbyá populations from the Brazilian Atlantic coast (see panels C and D in the figure below).

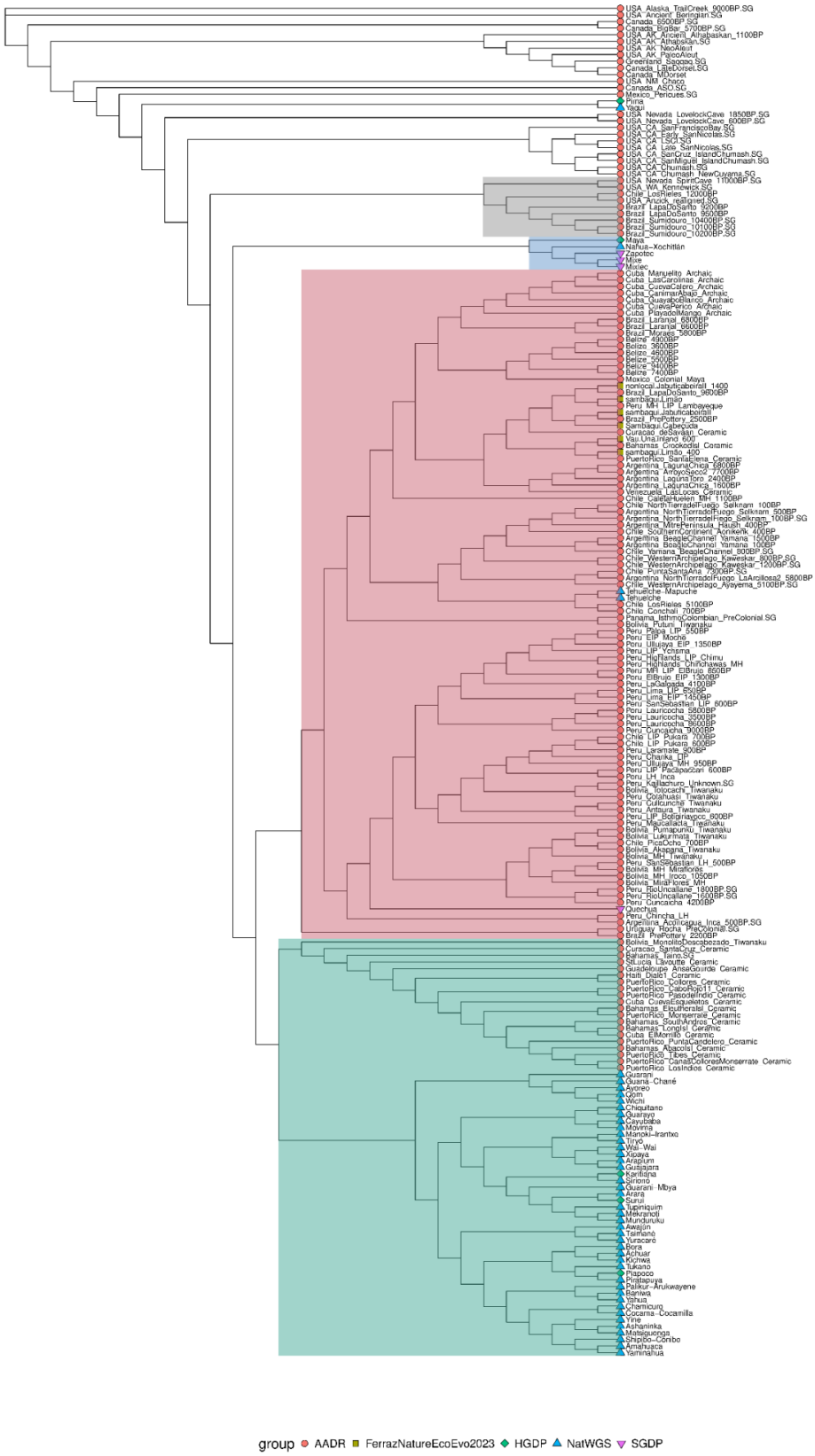

Newly sequenced data
Publicly available data
Indigenous American ancestry (%)
60
70
80
90
100
A
B
C
D
E
F
PC1 1.76%
PC2 1.46%
PC3 1.21%
PC4 0.5%
Admixed_American
East_Asian
NatWGS
Tupiniquim
South_Asian
African
European
Native_American
Northern_Asian_Siberian
Central_South_Asian
Middle_Eastern
Guarani-Mbya
Oceanian
1000_Genomes
HGDP
SGDP
NatWGS

In analyses including both present-day and ancient individuals, the data generation methods differ: present-day genomes were obtained exclusively through whole-genome sequencing, while ancient genomes were derived using either the 1240k SNP capture platform or whole-genome sequencing. Even in this context, we observed no evidence of batch effects, as individuals do not cluster according to the sequencing/genotyping methods or platforms. This can be verified in the NJ tree ($1 - f_3$(Mbuti; X, Y) presented in Figure 4A, where individuals are color-coded by data generation method (see figure below; please note the clades encompassing populations from the three proposed dispersals are highlighted by blue, red, and green rectangles):

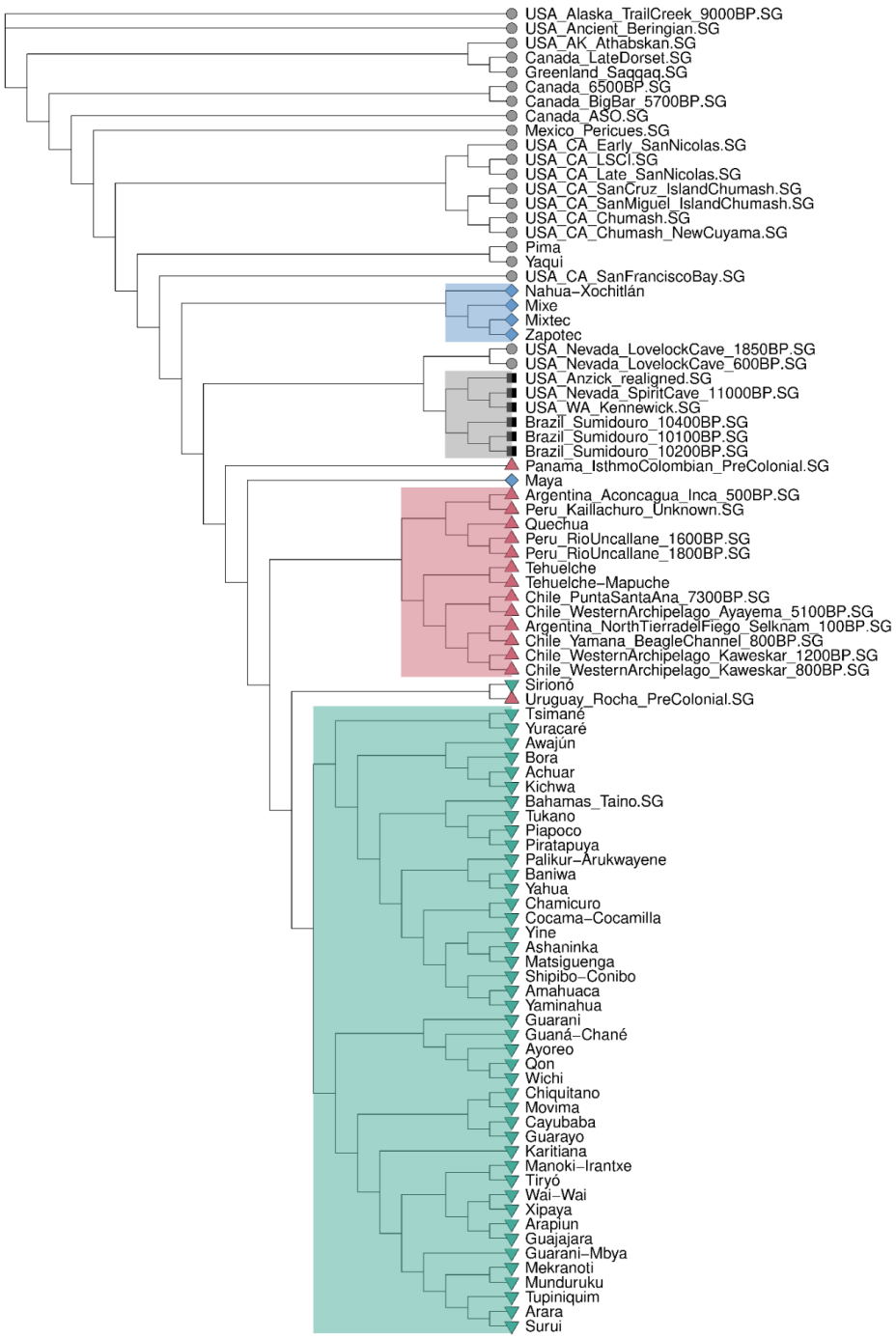


group .. 1240K BGISEQ-500 Illumina HiSeq X Illumina NovaSeq 6000 Shotgun

The absence of batch effects in the clustering pattern is further verified when populations are color-coded according to their source dataset.



Furthermore, when we repeated the analysis presented in Figure 4A using only the subset of data generated through whole-genome sequencing, thus excluding datasets generated through 1240K SNP capture, the results very closely mirrored those obtained from the full dataset (see figure below, with clades corresponding to the three proposed dispersals again highlighted by colored rectangles; additionally in blue we show the present-day Mesoamericans). This concordance indicates that the main findings remain robust even when the analysis is restricted to this whole-genome sequencing subset.



Regarding the possibility that the observed genetic similarity between most present-day Indigenous South American populations and Ceramic Period ancient Caribbean individuals is due to residual structure from earlier dispersals, this hypothesis also appears unlikely for a few reasons. It is improbable that such a residual structure would be present exclusively in the most recent individuals from the Caribbean (i.e., those from the Ceramic Period), while being entirely absent in individuals from earlier periods (prior to ~2,000 years ago) or from other regions such as the Andes or the Southern Cone. This distinct temporal and geographic pattern is inconsistent with what would be expected if the observed similarity were merely the result of ancient genetic structure dating back to around 9,000 years ago, when individuals associated with the second dispersal are first identified (or even at some point in time later). If that were the case, we would expect to see similar genetic affinities across a broader spatiotemporal range, which is not observed in the data.

Moreover, the patterns of genetic similarity observed when analyzing only the whole-genome sequencing data from present-day Indigenous Americans almost perfectly mirror those obtained when the analysis is restricted to the variants included in the 1240K array used in the joint analysis with ancient individuals. This consistency across analyses based on different marker sets underscores the robustness of the observed genetic similarity patterns, despite the more limited genetic resolution of the SNP set from the 1240K array SNP panel. It also supports the interpretation that the genetic structure present among contemporary South American lowland populations aligns with a broader, higher-level grouping (referred to here as the third-dispersal group), which also includes the Ceramic Period Caribbean individuals. Also reinforcing that this group is genetically distinct from earlier ancient populations or coeval ancient individuals from other regions of South America.

Additional strong evidence for the genetic differentiation and a distinct origin/population history of second- and third-dispersal populations comes from the automated admixture graph modelling approach that we have used. Here, we have selected one representative group from each of the genetic clusters highlighted in Figure 4A (since they would represent the most genetically differentiated groups included in our analysis) to infer the population history of these groups. As discussed in the text, the inferred admixture graph models evidence a genetic differentiation between second- and third-dispersal groups, which are consistently placed in independent branches of the population genealogies, evidencing a distinct origin and history of these groups. Most importantly, these analyses infer the requirement of substantive gene flow from Mesoamerican populations to explain the genetic diversity observed in third-dispersal groups. Taken together, these inferences (Figure 5) underscore the distinct population history of groups attributed to these distinct dispersals, corroborating the patterns of genetic similarity inferred before (Figure 4).

In response to the reviewer's comment 9 (please refer to the response for that comment for more details), we have also added a formal test to determine the minimum number of independent ancestry sources required to explain the genetic diversity observed in third-dispersal groups. Our analysis shows that models with rank 0 and 1 are strongly rejected ($p = 0$ and $p = 1.8 \times 10^{-25}$), while a rank 2 model is not rejected ($p = 1$; Supplementary Table 10).

This evidence shows that at least three distinct ancestry streams from ancient Indigenous Americans are needed to account for the observed genetic diversity in third-dispersal populations.

In the original manuscript we already applied the qpAdm framework to infer ancestry sources and their proportions, along with their standard errors, supported by statistical model fit p-values. This analysis showed that most present-day South American populations derive a significant portion of their ancestry from a source related to the North American Pacific Coast (Extended Data Figure 11A; Supplementary Table 10). In our admixture graph (Figure 5A), this source appears as a sister lineage to both present-day Mesoamericans (Mixe) and ancient individuals from the North American Pacific Coast. Consequently, the inferred contribution from the North American Pacific Coast reflects the gene flow from Mesoamerica into South America, an event we propose played a central role in shaping the genetic profile of the third-dispersal populations. The remaining ancestry derives from early SNA groups (first dispersal) and, to a lesser extent, second-dispersal groups, especially ancient Eastern South Americans. This admixture pattern is consistently supported across alternative top-ranked models (Extended Data Figure 11B; Supplementary Table 10).

Taken together, these multiple lines of evidence (absence of batch effects, consistency across WGS and 1240k datasets, replication in PCA, admixture graphs, qpAdm, and qpWave) strongly support the interpretation that the affinity between present-day South Americans and Ceramic Period Caribbean individuals reflects a genuine genetic connection, and not an artifact of data generation or residual structure.

Comment 8. Figure 4, which underpins these claims, suffers from overlapping panels (A and B) and missing branches. Labels in the map portion (panel B) are too small and overlap with icons. Panel C's axes and bar plots are difficult to interpret. In Figure 5, the "summary model" is unclear regarding what the "expected statistics" are, and the statistical test used to declare "no significant deviation" is not described.

Authors' reply: We thank the reviewer for their suggestions regarding the improvement of our figures and the presentation of results. In response, we have redesigned Figure 4 to avoid overplotting and have added more descriptive labels to panel C (see our response to comment 19 for further details). We have also provided high-resolution versions to ensure all details are clearly visible. This includes branches that were previously not visible due to the reduced resolution of the figures following their insertion into the text document.

Additionally, we respectfully disagree with the reviewer's interpretation that our conclusion regarding the existence of a third dispersal is underpinned on the results presented in Figure 4A. As outlined in our response to comment 7, this conclusion is supported by multiple lines of evidence. Figure 4 displays genetic similarity patterns based on a measure of genetic distance that is not affected by population-specific genetic drift, with further confirmation from the same analysis applied to present-day samples using WGS data (Figure 3A). Additionally, Extended

Data Figure 10 and Figure 5 present automated admixture graph modeling and a summary model incorporating selected representatives of the most genetically differentiated groups, further reinforcing our conclusion. Finally, qpWave (please refer to the response for that comment for more details) reveals that at least three distinct ancestry sources are necessary to explain the genetic diversity of third-dispersal populations, while qpAdm analysis infers the most likely sources and their ancestry contributions for each present-day Indigenous American population.

Regarding the results presented in Figure 5A, the main criterion for assessing an admixture graph model is the consistency between observed and predicted f-statistics. After fitting the model, residuals between observed and predicted f2, f3, and f4 statistics are analyzed. According to commonly accepted standards in the literature, any deviation with a |Z-score| of 3 or higher (that is, three or more standard deviations) across all statistics is generally considered significant. We have now clarified this statistical fit criterion more explicitly in the text.

> This summary model exhibited no significant deviations between observed and predicted f-statistics (maximum $|Z| < 3$; Figure 5A), although the complexity and number of parameters increased the potential for overfitting. However, when Quechua and Tehuelche individuals were included, the model fit deteriorated (maximum $|Z| = 4.43$), suggesting that their population history and ancestry profiles could not be adequately captured by this model or the sources used.

Comment 9. Overall, the interpretation of a novel dispersal would be more convincing if the authors performed formal tests (e.g., simulations) to show the data cannot be explained by simpler one- or two-wave models.

Authors' reply: We appreciate the reviewer's suggestion to strengthen the robustness of our findings. In response, we conducted a new analysis to formally assess the minimum number of ancestry sources required to explain the genetic diversity of third-dispersal populations. To achieve this, we employed qpWave, a formal statistical framework that tests the number of independent ancestry streams needed to model the demographic history of a set of populations using empirical allele frequency data. To do so, we incorporated all first- and second-dispersal populations as well as ancient individuals from the North American Pacific Coast as potential source populations (right populations).

The results indicate that both rank 0 and rank 1 models are strongly rejected ($p = 0$ and $p = 1.8 \times 10^{-25}$, respectively), while rank 2 is not ($p = 1$). This demonstrates that models assuming one or two ancestry streams are insufficient to account for the genetic diversity observed in the third-dispersal populations (i.e., left populations). Therefore, at least three distinct dispersal events involving groups with different ancestries are required, as rank 2 is the lowest rank not rejected.

Beyond that, in the original manuscript we already had employed the formal qpAdm framework to model the ancestry sources and estimate their admixture proportions, along with standard errors. Importantly, qpAdm provides a statistical model fit p-value that enables formal interpretation: high p-values indicate that the model cannot be rejected and, therefore, is considered plausible.

However, we acknowledge that the accuracy of the model depends heavily on the specific choices of outgroups and source populations. To address this limitation, as described in the Methods section and Supplementary Note 6, we applied a rotating approach using the qpadm_rotate function from the admixtools R package. This strategy allowed us to systematically explore a wide array of plausible ancestry models and evaluate their fit to present-day Indigenous American populations.

Specifically, we conducted two sets of analyses:

1) Excluding third-dispersal sources (restricted to ancient individuals): We restricted the sources to individuals associated with the first and second dispersals, as well as ancient individuals from the North American Pacific Coast, excluding present-day Indigenous Americans and Ceramic period ancient Caribbean individuals. This analysis tested 218 models per target population.

2) Complete set of sources (including both present-day and ancient individuals): We included representatives from all genetic clusters as potential sources and tested 5,811 models per target population.

For both analyses, we used a fixed set of outgroups (rightfix populations), which included: Mbuti (contemporary African group), French (representative of West Eurasian ancestry), and an ancient Beringian individual (USA_Ancient_Beringian.SG). In addition, all other populations were rotated as potential sources and outgroups (leftright populations), allowing us to test all plausible combinations for each target group. A summary of the top 50% of plausible models for each target population is presented in Extended Data Figure 11. The complete results, including standard errors for all models tested, are provided in Supplementary Table 10.

As described in the main text, our analysis excluding third-dispersal sources reveals that, for the majority of target populations (i.e., present-day populations for which we infer both ancestral sources and their proportional ancestry contributions), there is a consistent signal of ancestry from the North American Pacific Coast (purple bars in Extended Data Figure 11A).

In our summary admixture graph model (Figure 5A), this North American Pacific Coast source is modeled as a sister branch to present-day Mesoamericans, represented by the Mixe population. Therefore, the inferred ancestry contribution from the North American Pacific Coast reflects the Mesoamerican gene flow into South America that we propose played a central role in shaping the genetic profile of the third-dispersal populations.

The remaining ancestry contributions modelled by qpAdm are inferred to have originated from early SNA first-dispersal populations, and, less frequently, they also reflect contributions from second-dispersal groups, particularly ancient Eastern South Americans.

Therefore, combined analyses from multiple lines of evidence, including qpAdm results (Extended Data Figure 11), admixture graph modeling (Figure 5A and Extended Data Figure 10), outgroup f3 statistics (Figure 5B), and EEMS analysis (Extended Data Figure 4), support the conclusion that gene flow from North American populations (from populations closely related to ancient individuals from the North American Pacific Coast or present-day Mesoamerica) is necessary to explain both the genetic diversity found among present-day Indigenous South American populations and their genetic affinities with present-day Indigenous Mesoamericans.

The Neighbor-Joining tree constructed from outgroup f3-derived genetic distances (Figure 4A) highlights a genetic distinction between present-day Indigenous South Americans and ancient Indigenous American populations. Notably, present-day individuals exhibit a closer pattern of genetic affinities only with ancient Caribbean individuals from the Ceramic Period, as evidenced by their clustering within the same branch of the NJ tree. In addition, our new qpWave analyses formally demonstrate that populations associated with the third major dispersal into South America require at least three ancestral sources, using as potential sources first- and second-dispersal populations, as well as ancient populations from the North American Pacific Coast (which are a proxy for Mesoamerican-related ancestry).

Accordingly, we have revised the “A novel major dispersal into South America” section to incorporate the new findings and to more clearly articulate the multiple lines of evidence supporting the conclusion that the genetic diversity of Indigenous South Americans was shaped by at least three major population dispersals.

> To formally test the minimum number of ancestral sources needed to explain the genetic diversity observed in third-dispersal populations we used qpWave[47], including all first- and second-dispersal populations, as well as ancient individuals from the North American Pacific Coast, as candidate sources (right populations; Methods). This framework allowed us to test how many independent streams of ancestry (i.e. dispersal events) are necessary to account for the genetic variation observed in third-dispersal populations (left populations). Our results show that models with rank 0 and rank 1 are strongly rejected ($p = 0$ and $p = 1.8 \times 10^{-25}$, respectively), whereas a rank 2 model is not rejected ($p = 1$; Supplementary Table 10). These findings indicate that models assuming only one or two ancestry streams are insufficient to account for the observed genetic diversity of third-dispersal populations. Therefore, at least three distinct dispersal events involving genetically differentiated source groups are required, as rank 2 is the lowest rank not rejected.

Using qpAdm[44], we sought to identify the most likely ancestral sources contributing to present-day Indigenous Americans, initially excluding candidate source populations associated with the third dispersal (Methods). Our results indicate that most South American populations derive a substantial proportion of their ancestry from a source related to the North American Pacific Coast (Extended Data Figure 11A; Supplementary Table 10). In our summary admixture graph model (Figure 5A), this North American Pacific Coast source is positioned as a sister lineage to present-day Mesoamericans, represented by the Mixe population. Consequently, the inferred contribution from the North American Pacific Coast reflects the gene flow from Mesoamerica into South America, an event we propose played a central role in shaping the genetic profile of the third-dispersal populations. The remaining ancestry components modeled by qpAdm are attributed to early SNA groups from the first dispersal and, to a lesser extent, to populations associated with the second dispersal, particularly ancient Eastern South Americans. This admixture pattern was consistently supported across alternative top-ranked models (Extended Data Figure 11A; Supplementary Table 10).

Including all cluster representatives as sources (Methods) reveals ancestry patterns consistent with the previously identified genetic structure and affinity trends (Figures 2–4). The majority of present-day Indigenous South American ancestry is inferred to derive their ancestry from third-dispersal sources (Extended Data Figure 11B; Supplementary Table 10). While traces of genetic continuity with the first and second dispersals are evident in a few cases, particularly from the Early SNA and Ancient Eastern South American clusters. The Tehuelche and Quechua populations stand out for their distinctive ancestry profiles, since they primarily trace their ancestry to Early SNA or a combination of Early SNA and North American Pacific coastal sources. Notably, this pattern is consistently observed across the set of alternative high-probability models (Extended Data Figure 11B; Supplementary Table 10).

Our findings reveal that the genetic landscape of present-day Indigenous South Americans and Ceramic-period Caribbean populations was substantially shaped by an unreported third major dispersal likely originating from Mesoamerican-related groups. This dispersal would have occurred during the Late Holocene, preceding the arrival of ceramic-associated populations in the Caribbean and dating back at least 1,300 years, the age of the earliest individuals identified as part of this dispersal. The distinctive ancestry and affinities of these populations (Figure 4; Extended Data Figure 8C–D), the inferred genetic contributions from Mesoamerican-related groups (Figure 5A, Extended Data Figures 10–11), and the higher-than-expected affinity between Mesoamerican and South American Indigenous groups (Extended Data Figure 5) further support this scenario. The increasing genetic affinity for Mesoamericans over time (Figure 5B), particularly during the Late Holocene, reinforces the role of this third dispersal in shaping South American and Caribbean genetic diversity.

We have updated both the Methods section and the Supplementary Information to provide a more detailed description of the methodology employed for the qpWave analysis.

Methods

We used the qpWave[44] method to estimate the minimum number of independent ancestral sources contributing to third-dispersal populations, and the rotating qpAdmDM approach to model the sources and their proportional ancestry contributions to present-day Indigenous South Americans (Supplementary Note 6), both from the admixtools R package. Analyses were restricted to transversions with a maximum per-site missing rate of 10%. Two complementary analyses were performed: (i) including representatives from all genetic clusters, and (ii) focusing on individuals from the first and second dispersals, alongside ancient genomes from the North American Pacific Coast (i.e., excluding present-day Indigenous Americans and Ceramic-period Caribbeans). We identified feasible models in which source contributions summed to 100%, and plotted their probabilities and admixture proportions (Extended Data Figure 11). Full model statistics are reported in Supplementary Table 10.

Supplementary Information

We used the qpWave method to estimate the minimum number of independent ancestral sources contributing to third-dispersal populations (left populations), incorporating all first- and second-dispersal groups, along with ancient individuals from the North American Pacific Coast, as potential sources (right populations).

To explore a broad range of ancestry models and quantify their contributions to present-day Indigenous American populations, we also applied the rotating qpADM approach via the qpadm_rotate function in the admixtools R package[22]. Two analytical frameworks were implemented:

(i) The first analysis evaluated 5,811 models per target population, incorporating representatives from all major genetic clusters: Brazil_Sumidouro_10400BP.SG, Peru_Lauricocha_8600BP, Belize_9400BP, Argentina_ArroyoSeco2_7700BP, Argentina_BeagleChannel_Yamana_1500BP, Brazil_PrePottery_2500BP, Cuba_GuayaboBlanco_Archaic, USA_CA_Early_SanNicolas.SG, Mixe, Amahuaca, Karitiana, Guaná-Chané, Bahamas_Taino.SG.

(ii) The second analysis assessed 218 models per target, excluding present-day Indigenous Americans and Ceramic-period ancient Caribbeans. It included only representatives of the first and second dispersals, as well as the North American Pacific Coast cluster: Brazil_Sumidouro_10400BP.SG, Peru_Lauricocha_8600BP, Belize_9400BP, Argentina_ArroyoSeco2_7700BP, Argentina_BeagleChannel_Yamana_1500BP, Brazil_PrePottery_2500BP, Cuba_GuayaboBlanco_Archaic, USA_CA_Early_SanNicolas.SG.

A fixed set of outgroups ('rightfix' populations) was used, including contemporary Mbuti and French, along with an ancient Beringian individual (USA_Ancient_Beringian.SG). Additionally, all sources were tested as potential outgroups ('leftright' populations), allowing for the evaluation of all possible combinations of 'leftright' populations as either sources or outgroups for each target. Feasible models for each population are reported in Supplementary Table 10.

Comment 10. The authors demonstrate some affinity between Indigenous Americans and both Australasian and archaic hominin DNA. However, the timing frameworks need clarity. For instance, sections discussing runs of homozygosity (ROH) are confusing: long ROH typically indicates recent inbreeding, whereas significant proportions of ROH suggest historical inbreeding. It is unclear how ROH overlaps are interpreted as evidence for gene flow events that occurred ~10,000 years ago.

Authors' reply: We appreciate the reviewer's thoughtful comment and acknowledge that the current phrasing may have led to confusion. To clarify, we do not interpret the presence of ROH hotspots as evidence of gene flow from Population Y. Rather, we hypothesize that populations or genomic regions exhibiting signals of long-term and recent isolation, and thus a higher frequency of ROH, are less likely to have suffered influence from admixture events in both the distant and recent past, which could have diluted the original genetic signal of ancient admixture with Population Y. As a result, we hypothesized that these populations or genomic regions are more likely to have retained haplotypes originating from this ancient admixture event with the Ypykuéra ancestry source, which is believed to have occurred in Beringia or earlier in the lineage of Indigenous American ancestors. However, our results do not support this hypothesis, as only ~6% of the positions with excess Australasian affinity coincided with ROH hotspots.

We have revised the referenced paragraph to make this point clearer:

As previously proposed, another possibility for the origin of this affinity to Australasians is that the Ypykuéra ancestry component could have been retained more significantly only among isolated communities with high levels of internal genetic similarity[14]. Populations and genomic regions showing evidence of long-term or recent isolation, and therefore a higher frequency of runs of homozygosity (ROH), are less likely to have been affected by admixture events in the past, which could have obscured or diluted the original genetic signal of ancient admixture with Population Y. However, this scenario was not supported by our data, which revealed a lack of correlation between excess genetic similarity with Australasians and the inbreeding coefficient (FROH) (Spearman's r = 0.2503; p = 0.1192). Furthermore, our analysis reveals that hotspots of ROHs, defined as genomic regions with a significantly higher-than-average ROH density (more than 3 standard deviations), exhibit minimal overlap with loci displaying significant affinity to Australasians (Supplementary Note 5). Specifically,

only ~6% of the positions with excess Australasian affinity coincided with ROH hotspots.

Comment 11. Maintaining a low level of "Ypykuéra" (Australasian-like) ancestry for 10,000 years might be entirely plausible under neutral evolution; simulations could clarify whether this persistence truly exceeds neutral expectations.

Authors' reply: We understand and agree with the reviewer's point; however, we do not conclude that all Ypykuéra ancestry was maintained by positive selection in these populations. Rather, we had the hypothesis that this ancestry might have been at least partially preserved due to positive selection. Our analyses provide evidence that at least some genomic regions with stronger signals of this ancestry are consistent with having undergone positive selection.

We have revised the text to moderate its initial hypothesis framing. Specifically, we now make clear that we aimed to test the hypothesis that Ypykuéra ancestry may have been at least partially, rather than entirely, maintained through positive selection. We further clarify that our results support this possibility in certain genomic regions, but not all. Therefore, we conclude that while positive selection may have contributed to the persistence of this ancestry, it cannot fully account for it:

> The inferred proportion of Ypykuéra ancestry in present-day Indigenous populations is consistently low, ranging from 1 to 3%[36,37], and mirrors the estimates inferred for the Sumidouro individual, dated approximately 10,400 years before the present[5]. The persistence of this ancestry at similar levels over millennia suggests that some haplotypes may have been maintained by natural selection. To investigate their potential functional and evolutionary significance, we employed two complementary approaches to identify genomic regions in modern Indigenous American groups exhibiting excess genetic affinity to present-day Australasians: a local ancestry deviation test and a genome-wide D-statistic applied in an overlapping sliding-window framework.
>
> [...]
>
> Taken together, these findings suggest that several genomic regions exhibiting significant allele sharing with present-day Australasians may have been targets of positive selection. The candidate genes identified within these regions are implicated in critical biological processes with potential impacts on the health and adaptive history of Indigenous American populations. Nonetheless, it is important to note that not all regions exhibiting Ypykuéra ancestry are necessarily under selection. While positive selection may have contributed to the maintenance of this ancestry at specific loci, its scattered presence across other parts of the genome may also reflect stochastic events driven by neutral evolutionary processes.

Comment 12. The use of PBS to identify genomic regions under selection that also show Australasian affinity should be explained more carefully. PBS is designed to detect high differentiation from closely related populations, so its combination with distant affinity signals is non-trivial and needs more explicit rationale.

Authors' reply: We appreciate the reviewer's comment and the opportunity to clarify our rationale regarding the use of PBS. In our study, PBS was not applied to detect Australasian affinity or differentiation from Australasian groups, as we recognize that the divergence time between Australasians and Indigenous Americans is deeper than the temporal resolution at which PBS can effectively detect selection signals.

To identify genomic regions under natural selection in Indigenous Americans that have an excess of Australasian affinity, we first applied two independent genome-wide approaches: (i) local ancestry deviation, inferred with RFMix using a reference panel of unadmixed Indigenous Americans, Sub-Saharan Africans, Western Europeans, and Australasian populations, followed by calculation of Z-values for deviations from the genome-wide mean in Australasian ancestry; and (ii) the D-statistic, D(Mbuti, Onge; Mixe, Indigenous Americans), applied in overlapping sliding windows to detect excess allele sharing with Australasian groups. Regions with Z-values >3 in both analyses were classified as Australasian-like candidate selection regions. PBS was then used to assess whether these candidate regions also exhibit population-specific differentiation relative to closely related groups. We computed PBS genome-wide (including all variants as well as the Australasian-like set), using Indigenous Americans as the focal population, Siberians as the sister group, and East Asians as the outgroup. We then tested whether Australasian-like variants were in the top 1% and 5% tails of the PBS genome-wide distribution, which would indicate locus-specific differentiation in Indigenous Americans rather than variation shared with Siberians or East Asians, and thus candidate targets of natural selection underlying local adaptation.

We have revised the text to make this rationale clearer:

And in Supplementary Note 08.

> To identify genomic regions candidate to natural selection with excess Australasian ancestry in contemporary Indigenous Americans, we combined two independent approaches: (i) local ancestry deviation, in which local ancestry was inferred with RFMix using a reference panel including unadmixed Indigenous Americans, Sub-Saharan Africans, Western Europeans, and Australasian populations. For each ancestry among the genome, we then calculated a Z-value as (average ancestry proportion at the site - average genome-wide ancestry proportion)/ standard deviation of the genome-wide ancestry proportion; (ii) D-statistic for D(Mbuti, Onge; Mixe, Indigenous Americans). We performed the analysis using 2.5MB sliding windows, with a step size of 0.5Mb, a minimum of 10 SNPs per window and no missing data (maxmiss = 0). Only unrelated and unadmixed individuals were included.

Genomic positions with Z-value >3 for both statistics (local ancestry deviation and D-statistic) were classified as Australasian-like candidate selection regions. This approach identified 46,205 putative alleles distributed across 24 genomic regions (with nearby regions grouped as a single unit) spanning 17 chromosomes (Supplementary Fig. 9).

To investigate whether these regions show population-specific differentiation consistent with natural selection in America, we applied the PBS approach. PBS estimates the amount of population-specific genetic differentiation in a three-population framework, in our case, with Indigenous Americans as the focal population, Siberians as the sister group, and East Asians as the outgroup. We estimated PBS values for all variants genome-wide and obtained the empirical distribution, which serves as a demographic baseline control. We then evaluated whether the variants previously identified as Australasian-like candidates fell within the top 1% and 5% tails of the genome-wide PBS distribution, which would indicate differentiation specific to Indigenous Americans rather than variation shared with Siberians or East Asians, and thus candidate targets of natural selection underlying local adaptation.

And also in the text:

To investigate their potential functional and evolutionary significance, we employed two complementary approaches to identify genomic candidate selection regions in modern Indigenous American groups exhibiting excess genetic affinity to present-day Australasians: a local ancestry deviation test and a genome-wide D-statistic applied in an overlapping sliding-window framework. Regions with Z-values >3 in both analyses were classified as Ypykuéra ancestry candidate selection regions. Subsequently, to verify whether these regions are under natural selection specific in the Americas, we performed genome-wide selection scan based on allele frequency differentiation using population branch statistics (PBS), and identify if Australasian-like ancestry candidate regions fall with the top 1% and 5% of genome-wide PBS distribution (Supplementary Note 8).

Comment 13. In discussing archaic introgression, the manuscript suggests that "we were unable to determine whether introgression signals were associated with Indigenous American or European ancestral components." This ambiguity leaves the main conclusions about adaptive introgression in question. Addressing this would require clearer partitioning of ancestry components in the analysis.

Authors' reply: We thank the reviewer for this comment. As also noted by Reviewer 3, there was a typo in this sentence: it should read "they were unable to" rather than "we were unable to." The sentence refers to previous studies that were not able to distinguish whether the introgression signals were associated with Indigenous American or European ancestry components. In contrast, in our study we used only masked data, so that only the Indigenous

American ancestry component was evaluated. Therefore, we are able to state that the introgression signal is specific to the Indigenous American component. We have clarified this point in the discussion, Methods section and in Supplementary Note 09, and we apologize for the confusion caused by the typo.

Main text:

> However, unlike in the present study, they were unable to determine whether introgression signals were associated with Indigenous American or European ancestral components.

Methods:

> To identify genomic regions exhibiting signals of archaic introgression in ancestry-masked Indigenous Americans, we used segments detected by the SPrime method.

Supplementary Note 9:

> Putative archaic alleles introgressed into Indigenous American genomes were identified using the method implemented in SPrime. This tool employs a heuristic scoring approach that detects regions of high linkage disequilibrium (LD) in the target population (e.g., Indigenous American) by comparing them against an unadmixed outgroup, specifically the Mbuti population from Africa. This comparison allows for the identification of candidate introgressed regions by distinguishing segments that exhibit extended LD patterns indicative of archaic ancestry. For these analyses, we used genomic data masked for Native American ancestry.

Minor comments:

Comment 14. On page 8, the paper discusses the Arapium individual's spoken language but does not integrate this into the broader genetic narrative. Explain how this linguistic note informs the study's conclusions.

Authors' reply:

The Arapium individual exhibits the highest levels of non-Indigenous American ancestry in the dataset and is also the only one who speaks a creole language that evolved through contact between an Indigenous language (specifically Tupi-Guarani) and Portuguese during the early colonial period. This observation aligns with the genetic data, as this individual's unique linguistic background corresponds with their elevated levels of European and African admixture.

We have made this point clearer in the revised version of the manuscript:

We detected a significant positive correlation between Indigenous American ancestry proportion and ROH number (Spearman's r = 0.69; p = 3.16e–22), while both African (r = –0.45; p = 1.97e–11) and European ancestries (r = –0.68; p = 4.20e–22) were negatively correlated with ROH number, underscoring the increased ROH levels of Indigenous Americans and corroborating previous studies[34]. In this sense, the Arapium individual from Northeastern Brazilian Amazon expectedly showed very low ROH levels owing to its admixed ancestry profile (26.92% European and 11.25% African). This individual speaks Nheengatu, a creole language derived from Tupi-Guarani that evolved through contact with Portuguese during the early colonial period. Historically used as a *lingua franca* in Brazil, Nheengatu is still spoken in parts of the Amazon. This linguistic background reflects a history of increased contact and admixture with non-Indigenous populations in the community from which this individual originates, consistent with their elevated levels of African and European ancestries.

Comment 15. The manuscript states that data will be available but provides no details. Specify the repository or timeline for data release.

Authors' reply: As outlined in the "*Data and Materials Availability"* section, the dataset will be deposited in the European Genome-Phenome Archive (EGA), a secure repository for genetic and health-related data. Researchers interested in accessing the data can submit an application through the EGA website. The data will be made available at the time of publication.

Additionally, as stated in the same section, variant-level information will be made publicly available through an online variant browser, which will be launched along with the publication of the manuscript.

Comment 16. The text consistently mentions "Indigenous Americans", but the figures toggle between "Native Americans" and "Indigenous Americans". Use consistent terminology throughout.

Authors' reply: We have standardized the terminology throughout the manuscript to refer exclusively to "Indigenous Americans."

Comment 17. Figures 2, 4, 5, and many Extended Data (6–13) have small fonts, overlapping elements, and unclear legends.

Authors' reply: We appreciate the reviewer's suggestion to enhance the visualization of our results. In response, we have increased the font size of figure elements that we identified as potentially difficult to read. However, we would like to point out that the previous difficulty in reading the text within the figures was likely due, at least in part, to the low resolution of the

images embedded in the manuscript. As noted earlier, we have now provided high-resolution versions of all figures for the reviewers' reference.

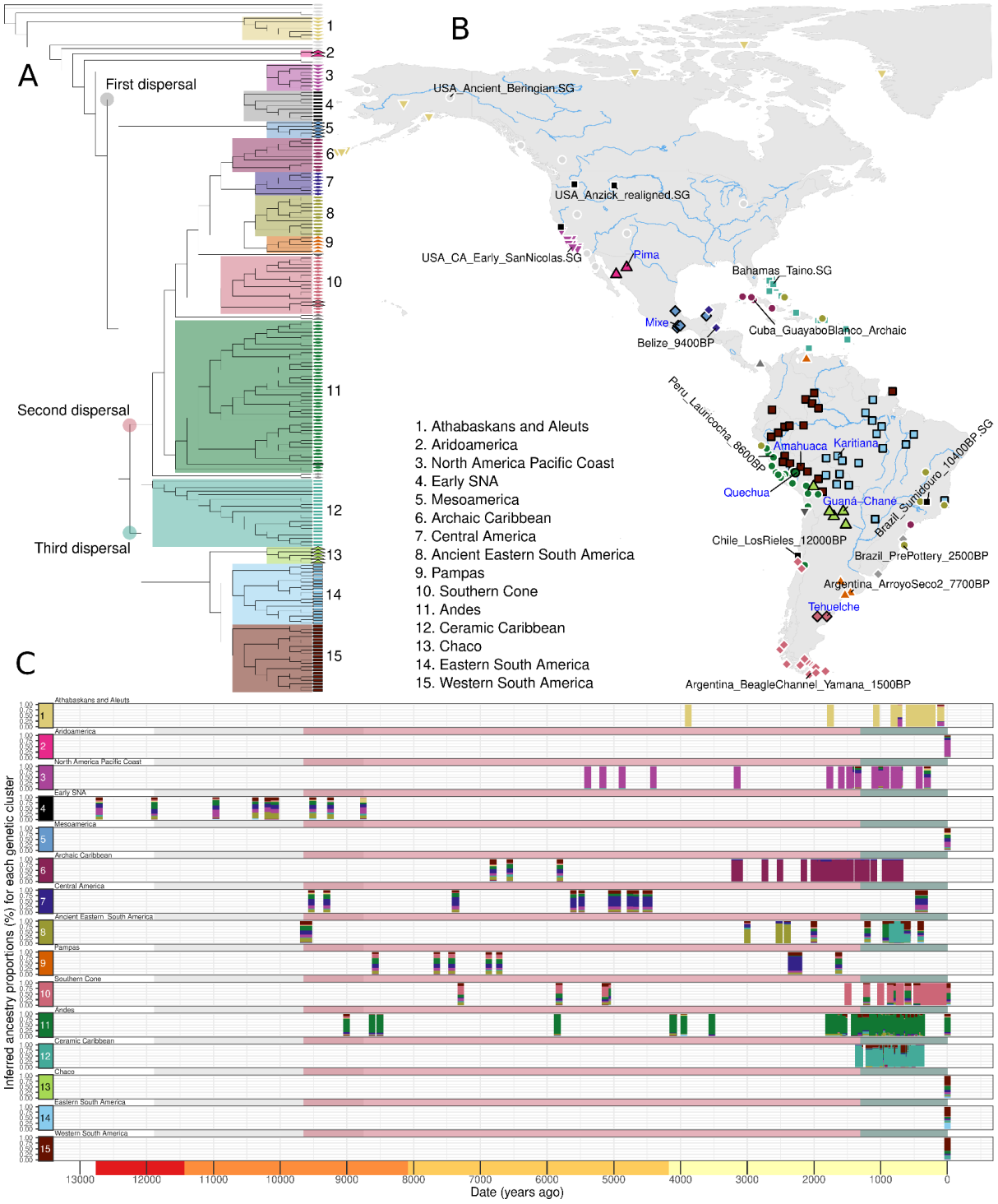
Comment 18. Figure 1 labels are in lowercase, whereas the text uses uppercase.

Authors' reply: Figure 1 panel labels have been updated to use uppercase letters, ensuring consistency with the formatting of panel labels in the other figures throughout the manuscript.

Comment 19. Figure 4: Panels A and B overlap; some branches are missing, and map icons are unreadable. Panel C's axes lack clear labeling.

Authors' reply: In response to the reviewer's suggestions, we have redesigned Figure 4 to reduce overplotting and added more descriptive labels to improve clarity and better link the information across panels. As mentioned previously, some figures were affected by low resolution and did not display correctly. To resolve this, we have now supplied high-resolution versions of all figures for the reviewers' reference.

The updated version of Figure 4 is also provided below (please note that the embedded version within this text and in the main text of the manuscript may suffer from resolution issues):



Comment 20. Provide explicit test statistics, effect sizes, or p-values when describing "unexpected" signals. Currently, the reader cannot confirm whether findings are robust or marginal.

Authors' reply: We thank the reviewer for highlighting this issue. We have now added explicit statements detailing the statistical tests performed, along with their corresponding effect sizes and significance levels.

Comment 21. Clarify if Arctic/North American populations were excluded fully or just from specific analyses.

Authors' reply: For our analyses, we have utilized all available data from the datasets included in this study, which comprise several present-day North American populations from Mexico, as well as numerous ancient individuals from Mexico, the USA, Canada, and Greenland. The only exception to this is the analysis of excess affinity between present-day and ancient Indigenous Americans and Australasians, for which we excluded Arctic and northern North American Indigenous groups. This exclusion was necessary because their ancestry stems partially or entirely from separate dispersal events from Siberia into the Americas. While this point is already addressed in the text (see highlighted sentence in blue), we now clarify that the exclusion of these groups applies specifically to this analysis (highlighted words in red).

> The second set of analyses identified at least one instance of significant excess genetic affinity in every cluster examined (Extended Data Figure 12; Supplementary Table 12), except for the Arctic and northern North American Indigenous groups, which were excluded specifically from this analysis because their ancestry was derived partially or entirely from independent dispersal events from Siberia into America[49–51] and, therefore, did not form a clade with other Indigenous Americans. The Sumidouro individual, dated approximately 10,400 years ago, remains the earliest known case of this excess affinity (Extended Data Figure 12). Notably, significant signals were detected from the Early Holocene to the present, with a marked increase in the number of individuals exhibiting this signal during the Late Holocene, particularly in the Andes, Pacific Coast, and western South America (Extended Data Figure 12). As previously noted, the apparently discontinuous spatiotemporal distribution of this ancestral component likely reflects the variation in its prevalence within and between populations[48]. These findings suggest that this ancestry was already present during the initial peopling of South America and may have played a more pronounced role in shaping the genetic landscape of the Late Holocene and contemporary populations.

Comment 22. Line 529: "bYpykuéra" should read "by Ypykuéra."

Authors' reply: As suggested, we have corrected this typo in the revised manuscript.

Comment 23. Line 1353: Add a space between the final paragraph of "Archaic introgression inference" and "Overrepresentation Analysis."

Authors' reply: As suggested, we have added the appropriate paragraph spacing.

Comment 24. Figure 5B: The x-axis label "date" has no units. Indicate if it is in years.

Authors' reply: In accordance with the reviewer's suggestion, we have updated the x-axis label of Figure 5B to include the necessary information.

REVIEWER 2

Dear Reviewer 2

We sincerely appreciate all of the reviewers' thoughtful comments, which we believe have significantly enhanced the quality and clarity of our manuscript.

Sincerely,

Tábita Hünemeier and colleagues

I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

REVIEWER 3

Dear Maria Ávila-Arcos

Thank you for your fruitful comments and valuable suggestions. Please find below your comments (marked in bold) and our replies following each one of them. The corresponding changes can be found in the manuscript attached with those marked in red.

Sincerely,

Tábita Hünemeier and colleagues

Overview:

I have finished reviewing the work "Indigenous American Genomes Harbor Unique Genetic Diversity and a Complex Evolutionary Past" by Castro e Silva and Nunes et al.

The manuscript describes the generation and analysis of the largest compendium to date of whole genomes from Indigenous Americans. This is a remarkable effort led and developed mainly by American (in the sense of "from The Americas" and not "from the USA") researchers. The study integrates the generated as well as previously available ancient and modern genomic data to gain numerous insights about the genetic variation and population history of the Indigenous peoples of the continent. They reveal the extent of new genetic variation observed in these genomes and its correlation with geography. They also show evidence and dating for three main dispersals into South America, signals of adaptive evolution, patterns of archaic introgression (adaptive and non-adaptive), and refinement of the genetic signal of Austronesian ancestry in some Indigenous Americans.

The amount of work and new insights is remarkable. This newly generated genomic data will advance many relevant studies on Indigenous populations of the Americas. I, however, find that the current version of the manuscript is very difficult to follow. The text, although generally well written, is extremely long, vague in some occasions and repetitive an intricate in others. A much-needed discussions on the ethical aspects, benefit sharing, and population engagement are lacking. The current structure and lack of important detail prevented me from making a more technical review (although I do have few technical comments). I hope that my comments help the methodological part be more transparent in a future version and then these other technical aspects can be more meticulously reviewed.

General comments:

Comment 1. Capitalize Indigenous throughout the text (sometimes it is and sometimes isn't).

Authors' reply: We would like to begin by thanking the reviewer for her helpful and insightful comments. We believe they have substantially improved the quality of our work and have contributed to making our results more accessible and valuable to both the research community and the broader public. We have capitalized the word "Indigenous" throughout the text in response.

Comment 2. The generated data does allow for some level of Structural Variation (SV) characterization. It is fair if the authors decided not to evaluate it, but it should be at least acknowledged that a lot of new variation could be potentially found by looking at SVs too.

Authors' reply: We appreciate the reviewer's insightful perspective and agree with the point raised. In line with this, we added the follow sentence in the new version of the manuscript:

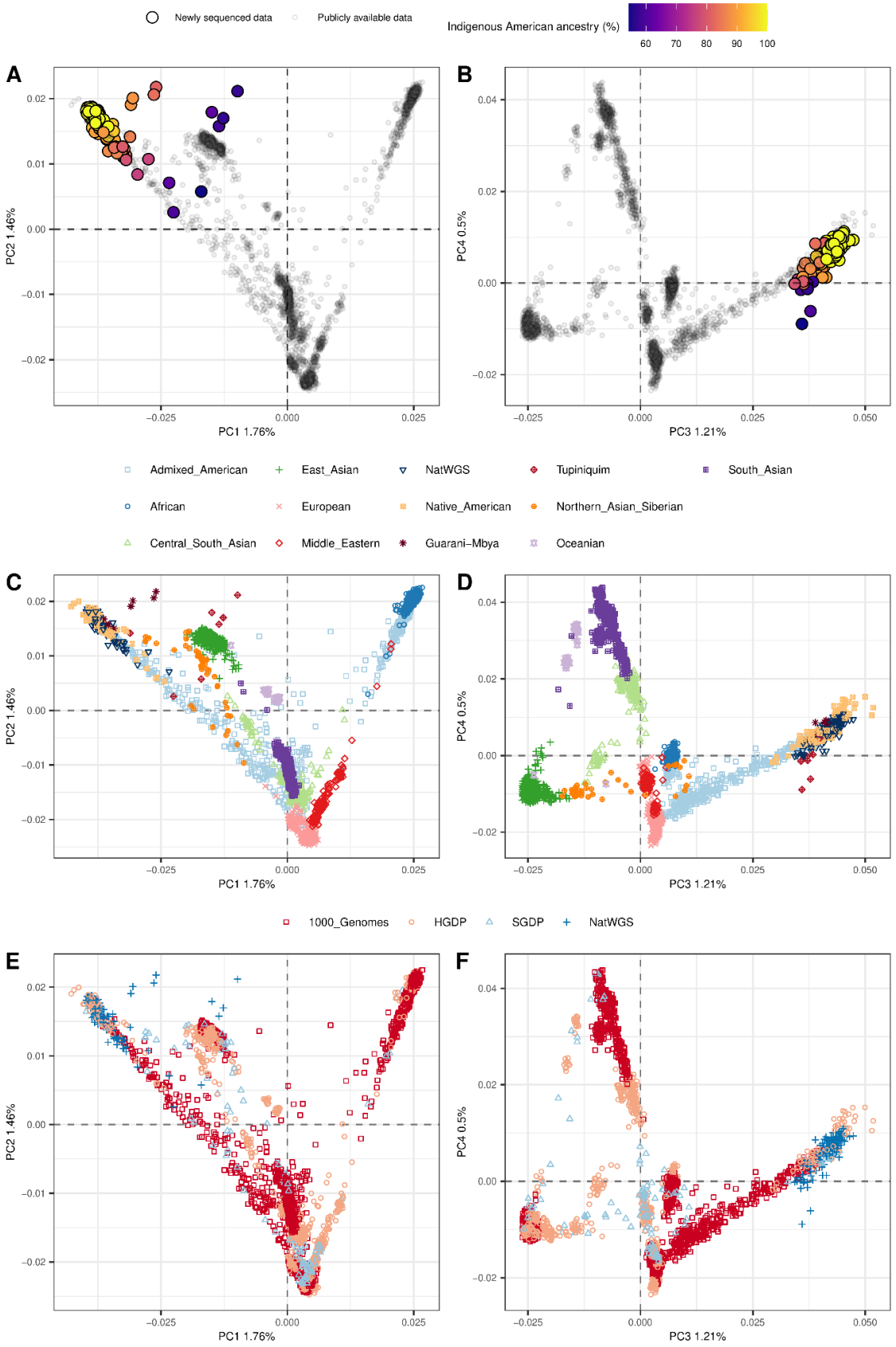
> Although we did not assess structural variation (SV) here, it is worth noting that these data could also uncover a substantial additional layer of novel genetic diversity at the SV level.

Comment 3. The generated genomic data comes from different sequencing platforms. It is known that generating data from different platforms can generate biases (e.g. making individuals sequenced or genotyped on the same platform artificially closer to each other). How did the authors check this didn't influence their analysis?

Authors' reply: We thank the reviewer for her comment. However, we formally tested for potential "batch effects," and we do not detect any signals in these data. First, as detailed in the Supplementary Material, the newly generated genomes were sequenced by two facilities: the Beijing Genomics Institute (BGI), using the BGISEQ-500 platform, and Dasa Genômica, using the Illumina NovaSeq 6000. Of the 128 newly sequenced individuals, 127 were sequenced at BGI, and only one individual was sequenced at Dasa. This distribution minimizes the potential for platform-induced artifacts and precludes artificial genetic affinities among the newly generated genomes due to sequencing batch, since only one individual comes from a different platform.

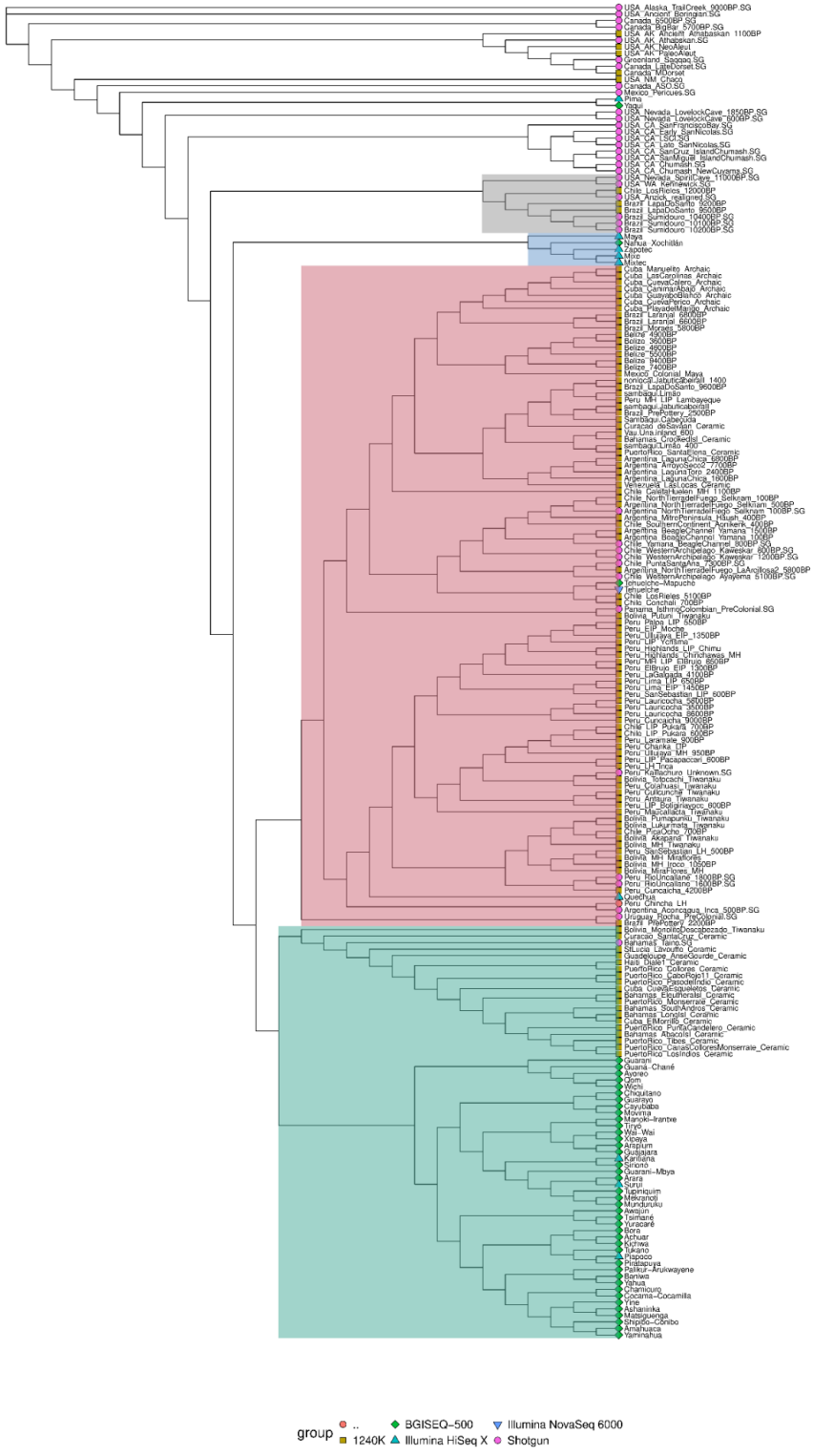
Second, we performed a PCA including all present-day individuals (presented in Supplementary Figure 2 in Supplementary Note 1), comprising both the 128 newly sequenced genomes and individuals from public datasets (1000 Genomes Project, Human Genome Diversity Project, and Simons Genome Diversity Project). As noted in the main text, *"PCA of global reference populations revealed no evidence of batch effects, as individuals were consistently clustered by geographic origin rather than by dataset source (e.g., HGDP, SGDP, or 1KGP), indicating robust integration across datasets (Supplementary Note 1)."*

In this analysis, the newly sequenced individuals (dark blue) cluster closely with publicly available Indigenous American genomes (light orange), supporting the absence of batch effects among Indigenous American individuals (Supplementary Figure 2). The few outliers observed correspond to individuals with higher proportions of non-Indigenous American ancestry (see panels A and B in the figure below), such as those from the Tupiniquim and Guarani-Mbyá populations from the Brazilian Atlantic coast (see panels C and D in the figure below).

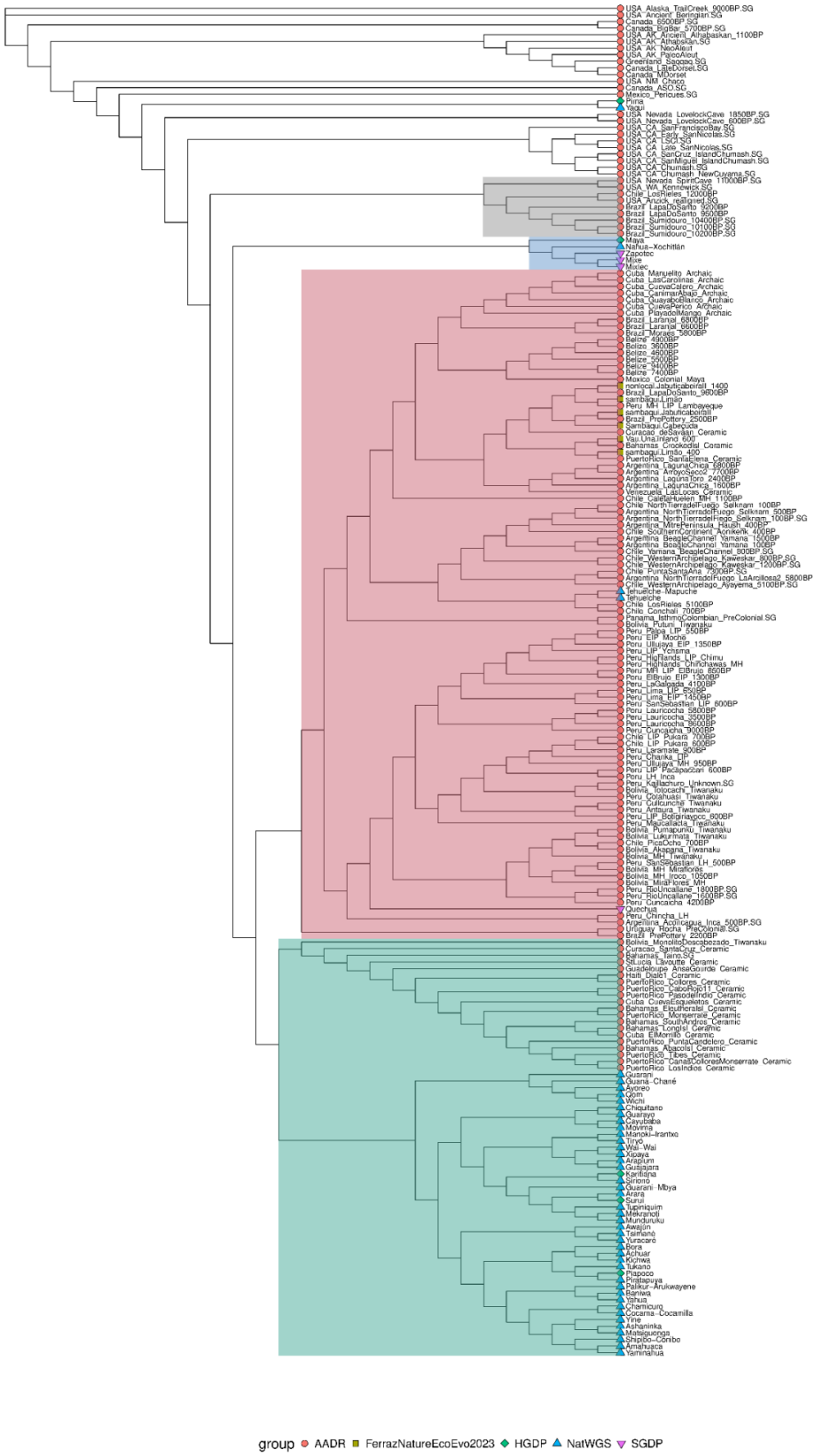


In analyses including both present-day and ancient individuals, the data generation methods differ: present-day genomes were obtained exclusively through whole-genome sequencing, while ancient genomes were derived using either the 1240k SNP capture platform or whole-

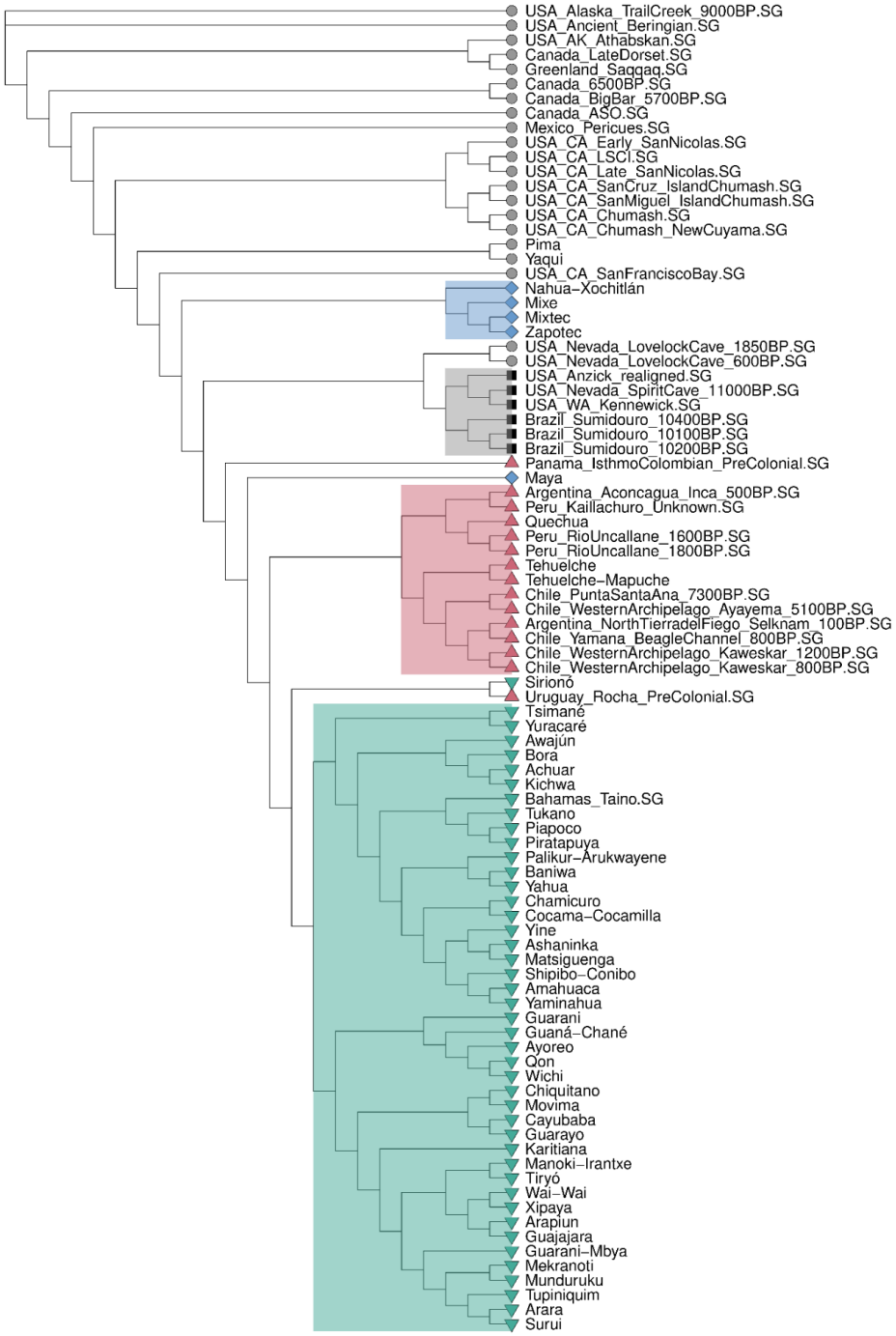
genome sequencing. Even in this context, we observed no evidence of batch effects, as individuals do not cluster according to the sequencing/genotyping methods or platforms. This can be verified in a plot of the results presented in Figure 4A, where individuals are color-coded by data generation method (see figure below; please note the clades encompassing populations from the three proposed dispersals are highlighted by colored rectangles):



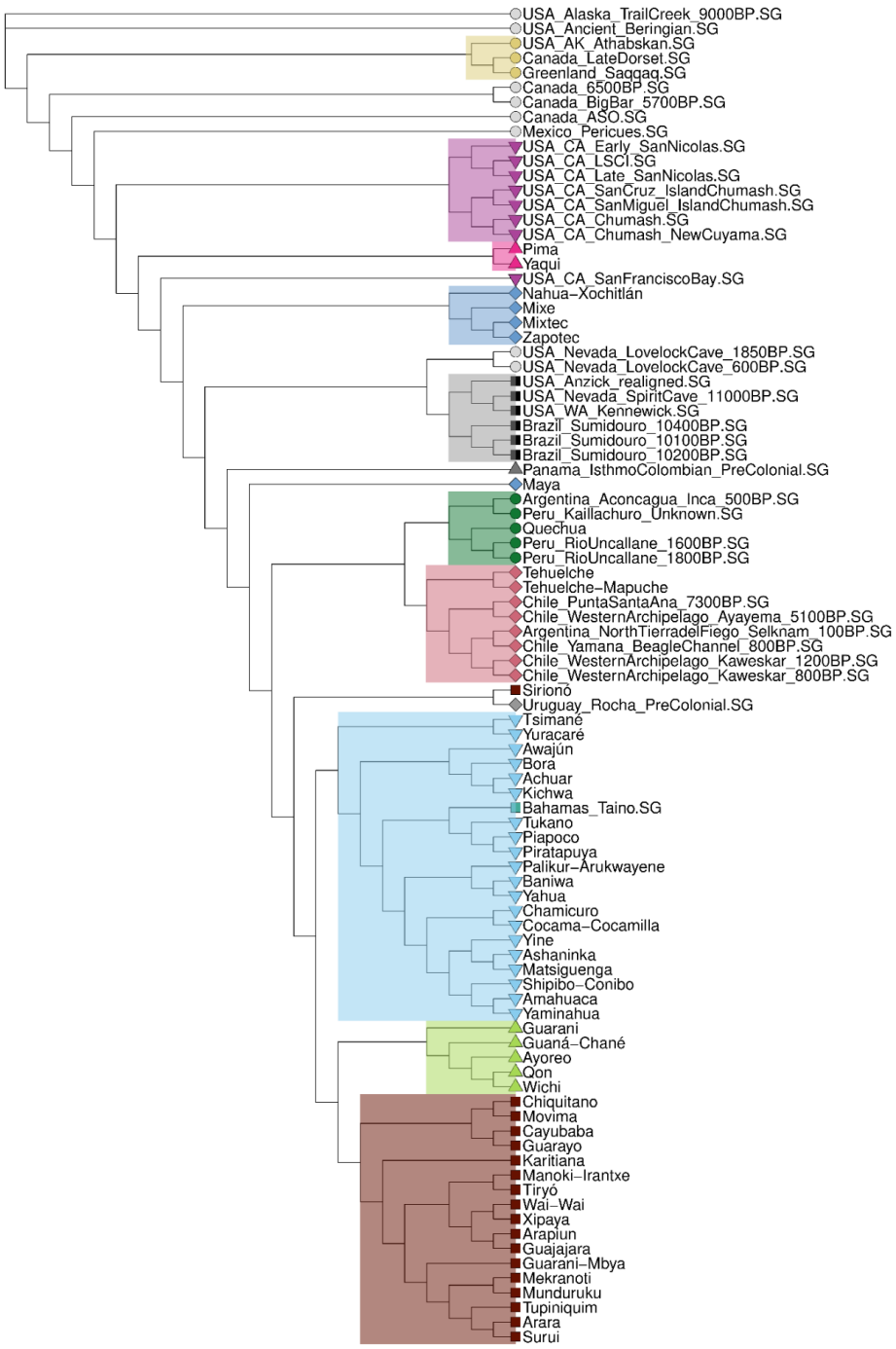
The absence of batch effects in the clustering pattern is further verified when populations are color-coded according to their source dataset.



Furthermore, when we repeated the analysis presented in Figure 4A using only the subset of data generated through whole-genome sequencing platforms, the results very closely mirrored those obtained from the full dataset (see figure below, with clades corresponding to the three proposed dispersals again highlighted by colored rectangles; along with present-day Mesoamericans in blue). This concordance indicates that the main findings remain robust even when the analysis is restricted to this WGS subset.



For the genetic clusters identified in Figure 4A, the clustering patterns are robustly reproduced when the analysis is restricted to WGS data, showing near-complete concordance between the two approaches (see figure below, where clusters are color-coded as in Figure 4A). Only a very small number of populations display discrepancies in cluster assignment relative to the full dataset shown in Figure 4A (tree tips are colored according to their original assignments from Figure 4A). Note that certain clusters are absent from the WGS-only analysis because no representative samples remain.



Comment 4. The main text and/or methods should give an overview of the ethical aspects of the sampling. This can be detailed in supplementary, but acknowledgment and reassurance that the work was culturally sensitive, ethical, and conducted following local regulations, should be made explicit early.

Authors' reply: We understand and agree with the concerns raised by the reviewer. In the original version of the manuscript, these topics were discussed in Supplementary Note 10, which we acknowledge was not the most appropriate placement. In the revised version, we have addressed this by incorporating a summary of the relevant content into the "Data Overview" section of the main text.

> All data included in this study were collected, generated and analyzed in strict accordance with international ethical standards for research with Indigenous populations (Supplementary Note 10). Our work was conducted in active dialogue and collaboration with Indigenous communities, and results were returned in accessible formats. Community representatives provided constructive and positive feedback, highlighting ways in which the genomic findings could complement and be integrated with their traditional knowledge and historical narratives.

> Supplementary Note 10:
> Ethical approval for sample collection was provided by the respective local ethics committees in each country: Argentina (Puerto Madryn Zonal Hospital, Resolution No 009/2015; San Carlos de Bariloche Zonal Hospital, Resolution No 1510/2015), Brazil (CONEP, Resolution No. 123, 763, and 4599), Bolivia (Universidad Mayor de San Andrés), Ecuador (Universidad de Las Américas; Consejo Nacional de Ciencia y Tecnología—CONACyT, grant # 69856; Instituto Nacional de Ciencias Médicas y de la Nutrición Salvador Zubirán Ref.: 1518), Mexico (Consejo Nacional de Ciencia y Tecnología—CONACyT, grant # 69856; Instituto Nacional de Ciencias Médicas y de la Nutrición Salvador Zubirán Ref.: 1518, CNIC Salud 2013-01-201471; Committee of Ethics and Research, UADY, official notice no. F-FENC-SAC-14/REV: 04; registry no. 09/17) and Peru (Universidad San Martín de Porres). Individual and/or tribal informed oral consent was obtained from participants who were unable to read or write. The ethics committees also approved the oral consent procedure. Logistical support for sample collection in Brazil was provided by the Fundação Nacional do Índio. All sampling was conducted in accordance with the Declaration of Helsinki and the applicable laws and regulations of each country at the time of sampling.
>
> The results of this study were presented to the participating communities. Results were returned to the communities in different ways, depending on the country to which each population belongs. Individual data remain anonymous; therefore, general results for each population were directly communicated to official representatives in Brazil and Argentina. Co-author P. S., a member of the Baré community, acted as a direct interlocutor with Indigenous populations in Brazil. In Mexico, the general results were

presented and discussed with the participants and students of junior high and high schools, and one of our coauthors (M.E.M.M.) is part of the Maya communities in which a long-standing collaborative project has been in place (coordinated by J.C.L.R.). In Bolívia, Peru and Ecuador, results were discussed with the populations through their national representatives. The data from this manuscript have been made available to community representatives. Where possible, clinical tests and medical advice were given to the participant communities (not restricted to participants only) and if needed, follow-up visits were performed to deliver results from both the main research and the clinical tests. In the future, a series of workshops will be organized, along with an accessible version of the manuscript, to be made available to the involved communities.

Comment 5. Ideally, a repository with documented scripts/commands used to analyze the data should be provided. This is key for reproducibility.

Authors' reply: We appreciate the reviewer's comment. Although our analyses were conducted using publicly available programs, with only a few minor exceptions, we will make relevant scripts available in a public repository prior to publication to ensure full reproducibility and facilitate access to our methodology.

Comment 6. A general difficulty when reading the text is the mixed use of terms to refer to populations: sometimes by language group, sometimes by genetic cluster, sometimes by population name or broad region (eg. Aridoamerica) and does not always follow an intuitive criteria. Something similar happens when referring to temporalities (e.g. Late Pleistoscene, Middle/Late Holocene, pre Ceramic, Archaic, etc). I mention some examples in the specfiic comments, but the text needs to be revised throughout with this in mind so the reader can understand better the geographic location of each.

Authors' reply: We thank the reviewer for her thoughtful perspective on this issue. It was indeed challenging to adopt consistent and clear terminology when referring to these groups throughout the manuscript. We have revised the text to better standardize the terminology and have provided definitions upon first mention to improve clarity.

This relationship between genetic similarity and geographic distribution becomes clearer through outgroup f3 statistics, which provide a measure of genetic similarity that is unbiased by population-specific genetic drift. A neighbor-joining (NJ) tree and multidimensional scaling (MDS) applied to 1–outgroup f_3 and 1/outgroup f_3 (Methods; Figure 3; Extended Data Figure 8A–B; Supplementary Table 5) revealed four major genetic clusters within South America that align closely with geographic regions. Accordingly, we named them based on their respective locations: (i) Southern Cone, (ii) Eastern, (iii) Western (mostly lowlands) South America, and (iv) Chaco. The Southern Cone populations (Mapuche-Tehuelche) were the most differentiated in South America, yet the Andean Quechua cluster was positioned within the same clade,

suggesting a long-distance genetic connection. The Chaco populations form the sister branch, followed by Tsimané, which remains unclustered, possibly because of their cultural and linguistic isolation[25,26]. The final division separates Eastern and Western South American lowland populations.

In North America, the first major split separates Pima and Yaqui into distinct branches (both in northern Mexico), followed by the divergence of the Nahua-Xochitlán group (central Mexico/Mexican Atlantic coast). Next, a cluster comprising the Mixe, Mixtec, and Zapotec branches off (on the southern Mexico), followed by the Maya (on the Mexican Yucatán Peninsula), and finally, the remaining continental diversity diverges. This pattern aligns with the traditional archaeological division between Aridoamerica (northern Mexico) and Mesoamerica, reflecting linguistic distinctions. The Pima and Yaqui belong to the Uto-Aztecan language family, whereas the Mesoamerican groups are associated with other language families. Based on this concordance between genetic, archaeological, and linguistic data, we group these populations into two genetically distinct clusters: Aridoamerica and Mesoamerica (Figure 3; Extended Data Figure 8A). While largely consistent with prior studies[11,27–32], the patterns of genetic structure observed here among Indigenous populations in North and South America reveal a more detailed substructure within these regions. These findings also strengthen the link between genetic differentiation and geographic distribution, supporting the trends identified in the PCA and ADMIXTURE analyses (Figure 2).

Regarding the use of geological terms (such as Pleistocene, Holocene, and Middle/Late Holocene) versus archaeological terms (such as the Archaic and Ceramic Periods) to describe temporal frameworks, we have now clarified our reasoning for choosing one timescale over the other in specific contexts. In short, the manuscript consistently employs the geological timescale in order to unify and simplify the presentation of results across populations from different regions of the continent, where archaeological periodization varies. This clarification has been included at the first mention of a geological period.

Effective population size (N_e) histories inferred from IBD analysis (Methods), grouping individuals by genetic cluster, revealed a widespread post-European contact bottleneck across all Indigenous American populations, followed by signs of recent genetic diversity recovery in Chaco and western South America (Figure 3D). Additionally, a pronounced pre-colonial population expansion of approximately an order of magnitude (from 10^4 to 10^5) was detected in both eastern and western South America (Figure 3D), consistent with previous evidence of demographic expansion during the Late Holocene[35,36]. Throughout the manuscript, we refer to the geological timescale to describe broad temporal intervals, aiming to unify and simplify the presentation of results across diverse regions of the continent, where archaeological chronologies differ. Mesoamerican populations maintained consistently high N_e values, although they were surpassed by those of eastern and western South America between ~1,500 and 500 years ago, and by western South America and the Chaco in the most recent period. In contrast, Aridoamerica exhibits persistently low N_e values and experienced

the most pronounced post-European contact bottleneck (Figure 3D), consistent with the high amount of long IBD segments shared within this cluster (Extended Data Figure 6).

Archaeological periods are mentioned only in the context of the Caribbean. This exception is due to the close correspondence between genetic similarity patterns and shifts in material culture in the region, as previously demonstrated in the literature and as we further illustrate in the manuscript. This clarification has been added where Caribbean archaeological periods are first introduced:

> Around 9,000 years ago, a new genetic lineage emerged and began to spread throughout Central America, reaching areas such as southern Mexico, Belize, and Panama. This lineage also appeared in the Caribbean, particularly among individuals from Cuba dating to the Archaic archaeological period (~5,000-2,500 years ago), and in northern South America, including what is now Venezuela. We reference the archaeological period for Caribbean individuals because genetic patterns in the region align closely with shifts in material culture, as previously demonstrated[34,45,46] and as we further corroborated. This second lineage continued expanding into the southernmost regions of South America. The Quechua in Peru and the Tehuelche in Argentina show genetic continuity with this lineage that persists to the present day. This second wave of dispersal, which at least partially replaced earlier populations, has been previously documented[5,6].

Comment 7. Why are the UpoA and UpopA2 not addressed at all in the text. Strangely, Mixe is the selected population for many f3 and D tests, when it is Mixe in fact the modern population with evidence of UpoA, but there is no acknowledgement of this or justification for selecting it as a representative of Mesomerica. I think how discussing how UpopA 1 and 2 play a role in their demographic reconstructions is important.

Authors' reply: We understand the reviewer's concern and acknowledge that these two unsampled populations were not explicitly mentioned or discussed in the text. However, UPopA is represented in the summary model shown in Figure 5A (labeled as PopA alongside the corresponding edge), as it is inferred to have contributed to the ancestry of present-day Mixe (Moreno-Mayar et al. 2018). Whereas, the population referred to as UPopA2 was modeled as an unsampled source contributing to the ancestry of an ancient individual from the Cañada de la Virgen site in Guanajuato, Mexico (Villa-Islas et al. 2023). Since this individual was not part of our analyses, UPopA2 could not be incorporated into our admixture graph model.

In our analyses, we aimed to integrate all currently known aspects of population history involving the Indigenous American individuals included in our dataset. To this end, we selected the Mixe as the representative Mesoamerican population, allowing us to account for the ancestry contribution of UPopA in our population history inferences, including the summary admixture graph model presented in Figure 5A, as well as in related analyses such as the

f3(Mbuti; Mixe, X) statistics shown in Figure 5B. Moreover, the Mixe were used as the representative Mesoamerican population in the original study that modeled the ancestry contribution from population Y into Indigenous South Americans (Skoglund et al. 2015). To maintain consistency with previous research, we have standardized the use of the Mixe in our analyses to make inferences comparable (using D statistics) regarding genetic affinity with both Australasians and archaic hominins, and previous studies.

We have revised the relevant section to explicitly state that the ancestry contributions of both unsampled populations A and Y have been modeled in our summary admixture graph (Figure 5A).

> We integrated our findings with the existing body of knowledge about population history involving the Indigenous American individuals to construct a comprehensive admixture graph model (Methods; Figure 5). In this summary model, we also incorporated the ancestry contribution of the unsampled population A (denoted as PopA in Figure 5A), which was modeled as a source contributing to the ancestry of present-day Mixe individuals[5]. Our results indicate that a population history in which this unsampled population contributed to the ancestors of both present-day Mesoamericans and ancient individuals from California provides a good fit to the genetic data. It is noteworthy that gene flow into the third-dispersal populations was not modeled using only the Mixe as proxies for the source. Rather, the model incorporates both the Mixe and ancient individuals from the California Channel Islands as a sister branch for the ancestral source contributing to third-dispersal populations.
>
> Additionally, we successfully modeled the ancestry contribution of the unsampled population Y (short for Ypikuéra, meaning "ancestor" in a Tupi language), which was inferred to account for the excess genetic affinity observed between certain Indigenous American groups and Australasians[45] (labeled as PopY in Figure 5A and further explored in the following section). This summary model exhibited no significant deviations between observed and predicted f-statistics (maximum $|Z| < 3$; Figure 5A), although the complexity and number of parameters increased the potential for overfitting. However, when Quechua and Tehuelche individuals were included, the model fit deteriorated (maximum $|Z| = 4.43$), suggesting that their population history and ancestry profiles could not be adequately captured by this model or the sources used.

We have also provided a clearer explanation of the rationale for selecting the Mixe as the representative Mesoamerican population in our analyses of genetic affinities with Australasians and archaic hominins.

> We also investigated the relationships between some genetic affinities to investigate whether the assumed Ypykuéra ancestry in Indigenous Americans could represent a shared ancestry between Australasians and Indigenous Americans, tracing back to archaic hominins (Neanderthals and/or Denisovans). In particular, we examined the

relationship between D(Mbuti, Onge; Mixe, X) and D(Mbuti, Neanderthal or Denisova; Mixe, X), where X stands for any Indigenous American group (Supplementary Table 13). The Mixe were used as a Mesoamerican reference population to ensure consistency across analyses and comparability with previous studies that reported excess affinity with Australasians[45,46]. No significant correlation was observed between genetic affinities for Australasians and Neanderthals (Spearman's r = –0.006, p = 0.971) or Australasians and Denisovans (Spearman's r = –0.1002, p = 0.5372). However, a strong positive correlation was found between the affinities for Neanderthals and Denisovans (Spearman's r = 0.6572, p = 7.2e–06), which is consistent with the expectation that these archaic ancestries were homogeneously distributed in the founding populations.

Comment 8. Another feature that makes this manuscript difficult to follow at times is that the narrative follows an analyses-per-analyses flow. A more intuitive and engaging way to present the findings would be to discuss by topic or question to be answered, then the findings along with ALL the analyses that support them.

Authors' reply: We appreciate the reviewer's suggestions. However, we respectfully disagree with the characterization that our manuscript follows an analysis-by-analysis structure. Instead, the manuscript is organized thematically, with each section addressing specific questions and aspects of the demographic and evolutionary history of Indigenous Americans:

A) Present-day genetic diversity, structure, and demographic history

(i) We begin by characterizing levels of genetic diversity, including the identification of novel variants specific to Indigenous Americans (section "Indigenous American genomes reveal higher-than-expected novel diversity").

(ii) This is followed by an exploration of patterns of genetic structure and similarity, along with the spatial distribution of genetic variation among present-day Indigenous populations (section "Indigenous American genetic variation mirrors geographic distribution").

(iii) We then examine patterns of homozygosity, identity-by-descent, and coalescence rates to reconstruct the demographic history of present-day populations, both in the recent and distant past (section "Uneven genetic diversity distribution and the lasting impact of colonization").

B) Genetic continuity, population movement, and an Indigenous American population history model

(iv) Next, we investigate genetic similarity between present-day and ancient individuals to assess signals of genetic continuity or evidence of population movements (section "Three main dispersals into South America").

(v) We then present a broader attempt to model population history through an integrative framework involving both present-day and ancient individuals, aiming to explain the observed genetic variation patterns (section "A novel major dispersal into South America").

C) Specific questions regarding the Indigenous American population history: ancient admixture and archaic introgression

(vi) This is followed by an analysis of a more specific topic within Indigenous American population history, focusing on Y population ancestry and archaic hominin introgression (section "Genetic affinity with Australasians and archaic hominins").

D) Positive selection, admixture-enabled adaptation, and adaptive introgression

(vii) Finally, we explore evidence of positive selection in Indigenous American genomes, including potential adaptive traits that may have been inherited through admixture with Population Y or archaic hominins (sections "Genomic adaptation in Indigenous Americans", "Natural selection and persistence of Ypykuéra ancestry in Indigenous Americans", and "Adaptive introgression in Indigenous Americans").

Comment 9. Please don't refer to individuals as "Samples" (E.g. Figure 1) , this objectivizes the human beings that donated their samples. This wording is particularly problematic for Indigenous peoples in the Americas, who have been historically marginalized.

Authors' reply: We completely agree with the reviewer and appreciate her concerns. In the revised version of the manuscript, we have consistently used the term "individuals" and removed the term "samples".

Specific comments:

Comment 10. Lines 203-205: More detail should be given (in Methods or Supplementary) about said probabilistic model. Enough detail to be able to reproduce the findings is needed.

Authors' reply: We thank the reviewer for this valuable suggestion. In response, we have exp anded the description of our probabilistic model in the Methods section to ensure sufficient de tail for reproducibility. Redacted These additional details now ensure that our approach can b e fully reproduced.

We have updated the methods sections as follows:

> We estimated the expected number of new variants using a simple model:

Redacted

Comment 11. 224 – FROH is introduced to the unfamiliar reader for the first time here, but it is unclear what the F before ROH means.

Authors' reply: We thank the reviewer for this comment. In our analyses, F_{ROH} is defined as the ROH-based inbreeding coefficient. We have added this description at the first mention of the term in the manuscript.

Comment 12. Line 229- What is meant by genomic-geographic structure?

Authors' reply:

We thank the reviewer for bringing this point to our attention. We agree that the current phrasing is vague and misleading. We have now revised the sentence to more clearly and accurately convey the intended meaning:

> This relationship between genetic similarity and geographic distribution becomes clearer through outgroup f3 statistics, which provide a measure of genetic similarity that is unbiased by population-specific genetic drift.

Comment 13. Lines 229-239. The first paragraph describes the four major South American clusters. Then the paragraph in 239 starts referring to Pima and Yaqui. Then Nahua-Xochitlán, Mixe … etc, without never mentioning that these are populations outside South America. Authors should not expect readers to know the exact geographic location of each population. Although the figures show location, having to go constantly from text to figure to makes sense of the findings is very tiresome and not engaging. Also, it is unclear to what analysis or figure the first sentence of the paragraph is referring to.

Authors' reply: We thank the reviewer for bringing this to our attention. We overlooked the need to make this information more accessible, and the reviewer is absolutely right in highlighting this issue. In response, we have revised the text to explicitly state the geographic locations of these communities. Please note that the referenced section has been substantially reorganized in response to comments from this and the other reviewers.

> This relationship between genetic similarity and geographic distribution becomes clearer through outgroup f3 statistics, which provide a measure of genetic similarity that is unbiased by population-specific genetic drift. A neighbor-joining (NJ) tree and multidimensional scaling (MDS) applied to 1–outgroup f_3 and 1/outgroup f_3 (Methods; Figure 3; Extended Data Figure 8A–B; Supplementary Table 5) revealed four major genetic clusters within South America that align closely with geographic regions. Accordingly, we named them based on their respective locations: (i) Southern Cone, (ii) Eastern, (iii) Western (mostly lowlands) South America, and (iv) Chaco. The Southern Cone populations (Mapuche-Tehuelche) were the most differentiated in South America, yet the Andean Quechua cluster was positioned within the same clade, suggesting a long-distance genetic connection. The Chaco populations form the sister branch, followed by Tsimané, which remains unclustered, possibly because of their cultural and linguistic isolation[25,26]. The final division separates Eastern and Western South American lowland populations.
>
> In North America, the first major split separates Pima and Yaqui into distinct branches (both in northern Mexico), followed by the divergence of the Nahua-Xochitlán group (central Mexico/Mexican Atlantic coast). Next, a cluster comprising the Mixe, Mixtec, and Zapotec branches off (on the southern Mexico), followed by the Maya (on the Mexican Yucatán Peninsula), and finally, the remaining continental diversity diverges. This pattern aligns with the traditional archaeological division between Aridoamerica

(northern Mexico) and Mesoamerica, reflecting linguistic distinctions. The Pima and Yaqui belong to the Uto-Aztecan language family, whereas the Mesoamerican groups are associated with other language families. Based on this concordance between genetic, archaeological, and linguistic data, we group these populations into two genetically distinct clusters: Aridoamerica and Mesoamerica (Figure 3; Extended Data Figure 8A). While largely consistent with prior studies[11,27–32], the patterns of genetic structure observed here among Indigenous populations in North and South America reveal a more detailed substructure within these regions. These findings also strengthen the link between genetic differentiation and geographic distribution, supporting the trends identified in the PCA and ADMIXTURE analyses (Figure 2).

Comment 14. 266. Why is the signal of geneflow between Meso and South America unexpected? It has been reported previously.

Authors' reply: We agree with the reviewer that a higher migration rate between Mesoamerica and South America is not unexpected, as prior studies have already provided evidence of gene flow between these regions. We have revised the sentence in question to reflect this point and have included citations to earlier research that has evidenced this gene flow.

Additionally, an unexpected signal of increased migration rate between Mesoamerica and South America indicates greater genetic similarity than geographic distances alone would predict. This observation aligns with previous findings indicating gene flow from North and Central American populations into South American groups[5,33].

Comment 15. 270 – The title of this section does not seem to describe its content. At least the "lasting impact of colonization" is not described.

Authors' reply: We would like to respectfully disagree with the reviewer's argument. In this section, we describe several patterns observed in present-day populations regarding the distribution of ROHs and IBD segments, which we attribute to the impact of the colonization process, a point that is clearly stated in the text (with examples provided below; most relevant sections in blue).

Compared to other continents, Indigenous Americans exhibit a distinctive length distribution of runs of homozygosity (ROH), with a higher average number of segments across all length categories, except for the shortest ones (< 2 cM), where Oceanians show a greater number (Extended Data Figure 5), indicating both strong ancient population bottlenecks and recent inbreeding. The Mesoamerican and Southern Cone populations showed the lowest ROH levels among Indigenous Americans (Extended Data Figs. 3 and 5), suggesting greater genetic diversity. In contrast, Moseten (Tsimané), Pano-Takana (Amahuaca and Yaminahua), Zamuco (Ayoreo), and Tupi (e.g., Sirionó, Suruí, and Karitiana) have the highest ROH counts and total length

(Extended Data Figs. 3, 5; Supplementary Table 4). These groups also exhibited an excess of long ROH segments (Extended Data Figure 5), indicating recent inbreeding likely caused by population collapse, fragmentation, and post-colonization isolation.

[...]

Additionally, identity by descent (IBD) analyses revealed that segments >8 cM were predominantly shared within populations (Extended Data Figure 6; Supplementary Table 8), suggesting limited gene flow since the colonial period. Therefore, these small and isolated populations likely experienced increased inbreeding during this period, leading to the high prevalence of long ROHs (Extended Data Figure 5).

Effective population size (Ne) histories inferred from IBD analysis (methods), grouping individuals by genetic cluster, revealed a widespread post-European contact bottleneck across all Indigenous American populations, followed by signs of recent genetic diversity recovery in Chaco and western South America (Figure 3D). Additionally, a pronounced pre-colonial population expansion of approximately an order of magnitude was detected in both eastern and western South America (Figure 3D), consistent with previous evidence of demographic expansion during the Late Holocene[27,28]. Throughout the manuscript, we refer to the geological timescale to describe broad temporal intervals, aiming to unify and simplify the presentation of results across diverse regions of the continent, where archaeological chronologies differ. Mesoamerican populations maintained consistently high Ne values, although they were surpassed by those of eastern and western South America between ~1500 and 500 years ago, and by western South America and the Chaco in the most recent period. In contrast, Aridoamerica exhibits persistently low Ne values and experienced the most pronounced post-European contact bottleneck (Figure 3D), consistent with the high amount of long IBD segments shared within this cluster (Extended Data Figure 6).

That said, we agree that the persistent consequences for present-day Indigenous American genetic diversity may not have been sufficiently emphasized. To address this, we have added brief concluding remarks at the end of the section to more clearly highlight this interpretation (new text in red).

Finally, we inferred the N_e history of genetic clusters based on within-population coalescence rates (see Methods) and found that over the past 10,000 years the effective population size of Indigenous Americans declined by approximately 40–90%, corroborating previous findings[37]. This analysis also revealed that historically, Indigenous North American populations exhibited slightly higher N_e than their South American counterparts (Figure 3E). However, in the last 1,000 years, this trend has reversed, with South American lowland populations (Western and Eastern South American clusters) displaying marginally higher N_e. This pattern aligns with the IBD-based N_e inferences, which indicate a recent population expansion in the region (Figure 3D–E). Notably, IBD-based methods are sensitive to recent demographic events (within

the last 50–100 generations), but lack resolution for deeper timescales due to the unreliable detection of short segments. In contrast, coalescent-based approaches are more powerful at capturing ancient demographic events but lose resolution in the recent past, where fewer lineages remain and thus fewer coalescent events provide information about population history. Therefore, integrating both improves the inferences across temporal scales. Taken together, these findings illustrate the complex demographic history of Indigenous Americans, from their divergence from Asian ancestors to their genetic differentiation across the American continent. They also shed light on the lasting impacts of the colonization process on the levels and distribution of genetic diversity in present-day populations.

Comment 16. 288-The mentioning of the Arapium individual (with no geographic context) feels very sudden and lacking of context.

Authors' reply: We mention the Arapium individual here because he represents an extreme outlier in terms of ROH levels (see Extended Data Figure 5C). Our aim was to draw attention to this outlier status and provide context for it. We now included details about the individual's geographic origin and elaborated on why he was expected to have a more admixed ancestry and consequently lower ROH levels, in response to a related comment from Reviewer #1.

Specifically, the Arapium individual shows the highest levels of non-Indigenous American ancestry in the dataset and is the only individual who speaks a creole language that emerged from contact between an Indigenous language (Tupi-Guarani) and Portuguese during the early colonial period. This linguistic background is consistent with the genetic data, as it reflects a history of greater contact and admixture with non-Indigenous populations, explaining the elevated levels of European and African ancestry observed in this individual.

We have made these points clearer in the revised version of the manuscript:

We detected a significant positive correlation between Indigenous American ancestry proportion and ROH number (Spearman's r = 0.69; p = 3.16e–22), while both African (r = –0.45; p = 1.97e–11) and European ancestries (r = –0.68; p = 4.20e–22) were negatively correlated with ROH number, underscoring the increased ROH levels of Indigenous Americans and corroborating previous studies[34]. In this sense, the Arapium individual from Northeastern Brazilian Amazon expectedly showed very low ROH levels owing to its admixed ancestry profile (26.92% European and 11.25% African). This individual speaks Nheengatu, a creole language derived from Tupi-Guarani that evolved through contact with Portuguese during the early colonial period. Historically used as a *lingua franca* in Brazil, Nheengatu is still spoken in parts of the Amazon. This linguistic background reflects a history of increased contact and admixture with non-Indigenous populations in the community from which this individual originates, consistent with their elevated levels of African and European ancestries.

Comment 17. 301-Define IBD here since it is the first time it's mentioned.

Authors' reply: We have carefully reviewed the manuscript and ensured that acronyms are defined upon their first mention.

Comment 18. 328-330 This statement needs a reference, is it a figure in the manuscript? Is it a previous paper? And what's based on?

Authors' reply: We now clarify that this statement refers to the ADMIXTURE results in Figure 4C, which show the emergence of region-specific genetic components beginning during the Middle Holocene, and more prominently during the Late Holocene. To improve clarity, we have added a direct reference to this figure in the text and replaced specific dates with a broader reference to the Late Holocene, which better reflects the timeframe discussed.

> Within the continent, genetic differentiation among Indigenous American clusters intensified during the Middle Holocene, becoming particularly pronounced over the Late Holocene (Figure 4C).

Comment 19. 358- The inclusion of ancient "individuals" is mentioned with very little detail. A general description is needed to understand the reach of the analyses that involve them. What type of data (SNP capture I assume from methods)? From what temporal/spatial range?

Authors' reply: We acknowledge that this information was missing from the main text. In an effort to reduce the overall word count, we minimized content in several sections of the manuscript. As a result, this information was in Supplementary Note 3 in the original version, as follows:

> This dataset combined the local ancestry masked dataset with ancient DNA from The Allen Ancient DNA Resource (AADR)[12] and Sambaqui mound builders from Brazil[13], retaining only common variants. The Sambaqui data were genotyped using the 1240k SNP capture platform[14], and the AADR dataset included both ancient individuals genotyped with 1240k SNPs and those with whole-genome data restricted to the 1240k sites. After filtering for high-quality genomes (excluding samples with >90% missing data), the final dataset included 355 ancient individuals from the Americas, comprising 703,252 genome-wide SNPs.

We have now added a brief description of this in the main text, along with a direct reference to Supplementary Note 3, to ensure the relevant information is accessible and clearly cited.

To explore shared ancestry in Indigenous populations across time and space, we estimated the genetic affinity among all Indigenous Americans, incorporating ancient genomes from the Allen Ancient DNA Resource (AADR)[42] and Sambaqui mound builders from Brazil[15] (see Methods, Supplementary Note 3, and Supplementary Table 9).

[...]

The NJ tree reveals evidence for at least three major dispersal events into South America. The earliest of these is represented by the Southern Native American (SNA) branch, which includes individuals older than approximately 9,000 years. These individuals cluster together in the NJ tree (Figure 4; Extended Data Figure 8C–D) and exhibit a mixture of multiple ancestry components (Figure 4; Extended Data Figure 9). This pattern likely reflects an early, undifferentiated SNA ancestry that subsequently diverged into distinct lineages, consistent with previous findings[5,6,43]. Representative individuals from this early dispersal include: Anzick-1 from Montana, USA (~12,700 years ago), associated with the Clovis culture[44]; one individual from the Spirit Cave site in Nevada, USA (~11,000 years ago)[5]; one individual from the Los Rieles site in Chile (~12,000 years ago)[6]; and individuals from the Sumidouro and Lapa do Santo sites in Lagoa Santa, Brazil, dated to over 9,000 years ago[5,6].

Around 9,000 years ago, a new genetic lineage emerged and began to spread throughout Central America, reaching areas such as southern Mexico, Belize, and Panama. This lineage also appeared in the Caribbean, particularly among individuals from Cuba dating to the Archaic archaeological period (~5,000-2,500 years ago), and in northern South America, including what is now Venezuela. We reference the archaeological period for Caribbean individuals because genetic patterns in the region align closely with shifts in material culture, as previously demonstrated[34,45,46] and as we further corroborated. This second lineage continued expanding into the southernmost regions of South America. The Quechua in Peru and the Tehuelche in Argentina show genetic continuity with this lineage that persists to the present day. This second wave of dispersal, which at least partially replaced earlier populations, has been previously documented[5,6].

Comment 20. 374-388 Starting with "Our discussion focuses on.." is a weird way to start a paragraph that is not part of the discussion. Ideally you start paragraphs stating what you want to discover and what analyses / dataset you use to do so, then give a general idea of how your analyses support the findings. Then you close with a brief explanation of what the findings entail in the big picture of the paper. One would expect in this paragraph that all tree dispersals are described in general terms. Then details on each can be given in the following lines. Instead, here only two are mentioned. This contrast with the expectation makes the next paragraph (390-401) very difficult to follow.

Authors' reply: We appreciate the reviewer's feedback and acknowledge that the original text lacked clarity. In response, we have revised the entire section titled "Three main dispersals into South America" to improve its readability and present the information more effectively. Please note that we have also incorporated relevant suggestions from the other reviewers.

> To explore shared ancestry in Indigenous populations across time and space, we estimated the genetic affinity among all Indigenous Americans, incorporating ancient genomes from the Allen Ancient DNA Resource (AADR)[42] and Sambaqui mound builders from Brazil[15] (see Methods, Supplementary Note 3, and Supplementary Table 9). An NJ tree based on genetic distances (unbiased by population-specific drift) revealed a strong correlation between genetic similarity and geography, with some regions showing long-term continuity (Figure 4; Extended Data Figure 8C, D). Clustering patterns aligned with the ADMIXTURE results (Figure 4; Extended Data Figures 8–9), and formed geographically restricted clades that suggested at least partial genetic continuity within and limited mobility between these regions over time. Supporting this, we inferred a moderate correlation between genetic and geographic distances (Spearman's r = 0.46; p ≅ 0) when considering all ancient and present-day Indigenous Americans (similar to the results obtained by including only present-day populations; Figure 3; Extended Data Figure 8A–B). Additionally, we observed a significant but weaker correlation between genetic distances and the age gap between individuals (Spearman's r = 0.29; p ≅ 0). This correlation between genetic distances and both geographic distances and temporal differences was even more pronounced in North America but weaker in South America and between individuals from the two regions (Supplementary Note 4).
>
> The NJ tree reveals evidence for at least three major dispersal events into South America. The earliest of these is represented by the Southern Native American (SNA) branch, which includes individuals older than approximately 9,000 years. These individuals cluster together in the NJ tree (Figure 4; Extended Data Figure 8C–D) and exhibit a mixture of multiple ancestry components (Figure 4; Extended Data Figure 9). This pattern likely reflects an early, undifferentiated SNA ancestry that subsequently diverged into distinct lineages, consistent with previous findings[5,6,43]. Representative individuals from this early dispersal include: Anzick-1 from Montana, USA (~12,700 years ago), associated with the Clovis culture[44]; one individual from the Spirit Cave site in Nevada, USA (~11,000 years ago)[5]; one individual from the Los Rieles site in Chile (~12,000 years ago)[6]; and individuals from the Sumidouro and Lapa do Santo sites in Lagoa Santa, Brazil, dated to over 9,000 years ago[5,6].
>
> Around 9,000 years ago, a new genetic lineage emerged and began to spread throughout Central America, reaching areas such as southern Mexico, Belize, and Panama. This lineage also appeared in the Caribbean, particularly among individuals from Cuba dating to the Archaic archaeological period (~5,000-2,500 years ago), and in northern South America, including what is now Venezuela. We reference the archaeological period for Caribbean individuals because genetic patterns in the region align closely

with shifts in material culture, as previously demonstrated[34,45,46] and as we further corroborated. This second lineage continued expanding into the southernmost regions of South America. The Quechua in Peru and the Tehuelche in Argentina show genetic continuity with this lineage that persists to the present day. This second wave of dispersal, which at least partially replaced earlier populations, has been previously documented[5,6].

Comment 21. 441- "An exception is Spirit Cave" -> Authors need to be careful when referring to an individual and give the necessary spatiotemporal context so readers can interpret the statement. Spirit Cave is a place not an individual. This is something that happens in other places in the text, please look for them and change accordingly.

Authors' reply: We agree with the reviewer's point and have revised the text accordingly to address this issue throughout the manuscript. To improve clarity and avoid confusion, we have defined each ancient individual at their first mention.

The NJ tree reveals evidence for at least three major dispersal events into South America. The earliest of these is represented by the Southern Native American (SNA) branch, which includes individuals older than approximately 9,000 years. These individuals cluster together in the NJ tree (Figure 4; Extended Data Figure 8C–D) and exhibit a mixture of multiple ancestry components (Figure 4; Extended Data Figure 9). This pattern likely reflects an early, undifferentiated SNA ancestry that subsequently diverged into distinct lineages, consistent with previous findings[5,6,43]. Representative individuals from this early dispersal include: Anzick-1 from Montana, USA (~12,700 years ago), associated with the Clovis culture[44]; one individual from the Spirit Cave site in Nevada, USA (~11,000 years ago)[5]; one individual from the Los Rieles site in Chile (~12,000 years ago)[6]; and individuals from the Sumidouro and Lapa do Santo sites in Lagoa Santa, Brazil, dated to over 9,000 years ago[5,6].

In particular, we have partially rewritten the following paragraph, beginning at line 374 of the original manuscript:

Admixture graph analyses consistently placed present-day Indigenous South Americans in a clade with Ceramic-period ancient Caribbeans (Extended Data Figure 10A–C), whereas Central and South American ancient individuals, including Archaic-period Caribbeans, formed a separate clade, indicating distinct dispersal. This pattern further reinforces the genetic affinities observed earlier (Figure 4; Extended Data Figures 8–9), where present-day Indigenous South Americans exhibited stronger affinities with Ceramic-period Caribbeans, while second-dispersal ancient individuals showed greater genetic similarity among themselves. An exception is the individual dated to 11,000 years ago from the Spirit Cave site, which, despite showing affinity with the first dispersal clade (Figure 4), appears as the earliest split in the second dispersal clade (Extended Data Figure 10A–C).

Comment 22. 528- The phrasing is confusing. What "to others" mean here?

Authors' reply:

We thank the reviewer for raising this point. Our intention was to state that certain Indigenous American populations exhibit significantly greater genetic similarity to contemporary Australasians than other Indigenous American populations do. We have now revised the text to clarify this point and improve the phrasing accordingly:

> Certain Indigenous Americans exhibit significantly deeper genetic afinity to contemporary Australasians than to other Indigenous American groups[45,46], challenging the notion of a single non-Arctic Indigenous American clade.

Comment 23. 609 – This is the first mention in the text of PBS, so it should be named in full here.

Authors' reply: We appreciate this observation. We have carefully reviewed the manuscript and ensured that the full names of all methods are provided upon their first mention.

Comment 24. 675 – Put the correct name for SPrime and add the reference. Some mentions of Software lack references. Please check that all the software is credited properly.

Authors' reply: We thank the reviewer for this comment. The text has been carefully revised and all software tools and credits have now been referenced appropriately.

Comment 25. 719 – We -> they?

Authors' reply: We have made the necessary corrections in the revised version of the manuscript.

> However, unlike in the present study, they were unable to determine whether introgression signals were associated with Indigenous American or European ancestral components.

Comment 26. Figure 4 – The tree did not display correctly on my screen.

Authors' reply: We apologize for the low resolution of the figures in the final PDF. It appears that the submission system applies a compression method that reduces the resolution and

overall quality of the images. The original figures were prepared in high resolution and do not exhibit these issues. To address this, we have now provided an additional PDF containing high-resolution versions of all figures for the reviewers' reference.

REVIEWER 4

Dear Reviewer 4,

Thank you for your helpful comments and suggestions to improve our manuscript. Please find below your comments (marked in bold) and our replies following each one of them. The corresponding changes can be found in the manuscript attached, with those marked in red.

Sincerely,

Tábita Hünemeier and colleagues

Overview:
The authors make a valuable and exciting contribution to understanding the genomic population history of South America and the entire Americas. The central element of the study is the presentation of 128 new, high-coverage Indigenous American genomes from 45 populations and 28 language families, spread across several Central and South American countries. These new data are an invaluable resource for a continent and its peoples that have been seriously understudied. When combined with previously published modern and ancient genomes, they thoroughly analyze the data to find new insights into how South America was populated, the processes shaping its population structure, and aspects of evolutionary adaptation. Their key findings include new knowledge about previously uncharacterized levels of genetic diversity in Indigenous South American populations, as well as support for models that suggest multiple independent dispersals into South America, followed by processes of differentiation, and the identification of candidate genes under selection linked to various traits.

The strength of this study lies in its thorough analysis of genomic data, utilizing state-of-the-art population genetic methods. Unique and novel insights—beyond the essential data provided—are demonstrated, for example, by the models that help better understand the "population-Y" or "Ypykuéra ancestry." Here, the authors present the most compelling analysis to date of this ancestry component linked to Australasian populations. Additionally, the detailed analysis of genomic diversity provides critical empirical data to challenge the old "low genetic diversity paradigm" for Indigenous American populations. For these parts of the study, as well as the other valuable analyses, I see no potential issues; the methodology is solid, and the interpretations, to the best of my ability to assess, are supported by the data.

Comment 1. My critique of this manuscript is minor, and much of it may come down to style, personal preference, rather than essential issues. However, some points should be clarified, toned down, or adapted for the publication context. I want to start with a common critique of many "Novel Genome Collection" papers: it feels like the paper is overloaded with analyses, trying to cover every possible aspect, which sometimes leads to redundancies and sections that are harder to read. All analyses presented are valuable, but the authors missed an opportunity to contextualize their findings with the rich, interdisciplinary sources available for interpretation. There's a heavy focus on ancestry-related models, and the evolutionary discussions seem somewhat like a secondary addition. What is the significance of those findings? Don't get me wrong—I think the observations are vital, but it would have been helpful if they had been discussed and placed within a broader context. The authors could have easily expanded this into 2-3 separate manuscripts. But again, this is a matter of style, and it certainly shouldn't prevent this work from being published.

Authors' reply: First, we would like to thank the reviewer for the comments and constructive criticism. We believe they have significantly improved our manuscript.

In response to the feedback from the reviewers, we have more carefully integrated our results into the existing body of evidence on the demographic and evolutionary history of Indigenous Americans.

The reviewer will note that the manuscript has undergone a substantial restructuring, which we believe has significantly enhanced its clarity and impact. We have also placed greater emphasis on articulating the broader significance of our findings, while explicitly acknowledging certain limitations in specific analyses. In addition, we outline key next steps needed to more comprehensively characterize the genetic history of Indigenous Americans, integrating all these points at appropriate locations throughout the text.

Comment 2. One aspect that should be addressed is the claim of "novelty" in some sections. While this study does a fantastic job in verifying several hypotheses previously discussed and by adding a new level of complexity to our understanding, it feels like prior work on South American population history is undervalued (including that of some of the co-authors). Examples include the discussion of how population structure was shaped in South America, as well as the EEMS analyses. This is indeed the most detailed study on that, but both aspects of isolation-by-distance and environmental and socio-cultural factors, which are considered significant in shaping population structure, are not new and have been discussed by other researchers before. Another example is (Line 267, and associated text parts) the observed increased migration rate between Mesoamerica and South America. Why is this unexpected? This has been discussed in the literature, and a recent ancient DNA study of Colombian ancestors has shown evidence of important periods of Mesoamerican gene flow into South America. I do not mention these examples to devalue the importance of the author's study in any way. It just seems like claims of novelty should be somewhat toned down, or the work of others that addressed similar aspects should be made more visible.

Authors' reply: We appreciate the reviewer's concern and have revised the text to address it. Specifically, we have moderated some of the original claims, added additional context and references to prior work, and clearly indicated which elements of our findings are novel. We also tried to highlight the significance of these new insights within the broader framework of existing research.

To clarify that the genetic structure patterns observed in present-day populations are consistent with previous findings, despite those studies being primarily based on SNP-array data, we have added a sentence explicitly stating this point.

> While largely consistent with prior studies[11,27–32], the patterns of genetic structure observed here among Indigenous populations in North and South America reveal a more detailed substructure within these regions. These findings also strengthen the link between genetic differentiation and geographic distribution, supporting the trends identified in the PCA and ADMIXTURE analyses (Figure 2).

Specifically regarding the EEMS analysis, we have added a citation to the only prior study known to us. Notably, that study reported very similar patterns within its area of focus, namely western South America.

> In line with this, the estimated effective migration surface (EEMS) models indicated regions with genetic differentiation greater than expected under an isolation-by-distance model, notably southeastern Peru, northern Bolivia, and southwestern Amazonia (Extended Data Figure 4). This pattern closely resembles findings from a previous study involving a more limited set of populations from western South America[29].

We also agree with the reviewer that a higher migration rate between Mesoamerica and South America is not unexpected, as prior studies have already provided evidence of gene flow between these regions. We have revised the sentence in question to reflect this point and have included citations to earlier research that has evidenced this gene flow.

> Additionally, an unexpected signal of increased migration rate between Mesoamerica and South America indicates greater genetic similarity than geographic distances alone would predict. This observation aligns with previous findings indicating gene flow from North and Central American populations into South American groups[5,33].

Comment 3. The detection and description of the "third wave" of gene flow into South America is exciting; however, it sometimes seems that the way it is described could be misleading or confusing for the reader. For example, starting at Line 411, the authors state that "Almost all present-day indigenous South Americans… share a distinctive genetic affinity…." But later in the paragraph, they mention the exceptions, including the Andean Quechua speakers. A similar issue appears in Line 434, where it sounds like Quechua speakers are not Indigenous South Americans. It's important that the authors more carefully emphasize the geospatial aspects of the "third wave" expansion and influence.

Authors' reply: We understand the reviewer's concern and agree that our current phrasing in these sections may have been misleading. We have revised the sentences mentioned by the reviewer and carefully reviewed the entire text to identify and correct similar issues. Please note that we have also incorporated relevant suggestions from the other reviewers.

Line 411

> Several South American Indigenous groups clustered with ancient individuals from the second dispersal. For instance, the Quechua aligned with ancient populations from the Andes, whereas the Mapuche-Tehuelche and Tehuelche clustered with ancient individuals from the Southern Cone (a region that encompasses Patagonia, the southern Andes, and the far south). In both cases, the observed genetic clustering indicates a degree of continuity, as these present-day groups inhabit the same regions as the ancient populations to which they show the greatest genetic affinity. As expected, Indigenous North American populations are the most genetically distinct from their Indigenous South American counterparts. Populations from Aridoamerica diverged earlier than all lineages that later expanded into South America, whereas Mesoamerican groups branched off after the initial dispersal lineages but prior to the formation of the second-dispersal clusters (Figure 4; Extended Data Figure 8C–D).

Comment 4. Regarding the Quechua speakers, in line 475, they mention that these groups received "…some gene flow from ancient Andean individuals." This could be misleading. Aren't the "ancient Andean individuals" the ancestors of the Quechua? Or do the authors mean that modern-day Quechua speakers received gene flow in addition to their ancestral DNA?

Authors' reply: We agree with the reviewer that the current phrasing is misleading and does not clearly present our results. Our findings indicate that the Quechua show some degree of genetic continuity with ancient Andean populations. However, in our admixture graph modeling, we infer that the majority of their ancestry derives from the same branch as third-dispersal populations (Extended Data Figure 10G), a pattern also observed in the Tehuelche (Extended Data Figure 10F). For this reason, we originally described the ancient Andean contribution as gene flow, but we now recognize that this wording may obscure the interpretation and fail to provide a clear understanding for the reader. We have revised the text to clarify this point.

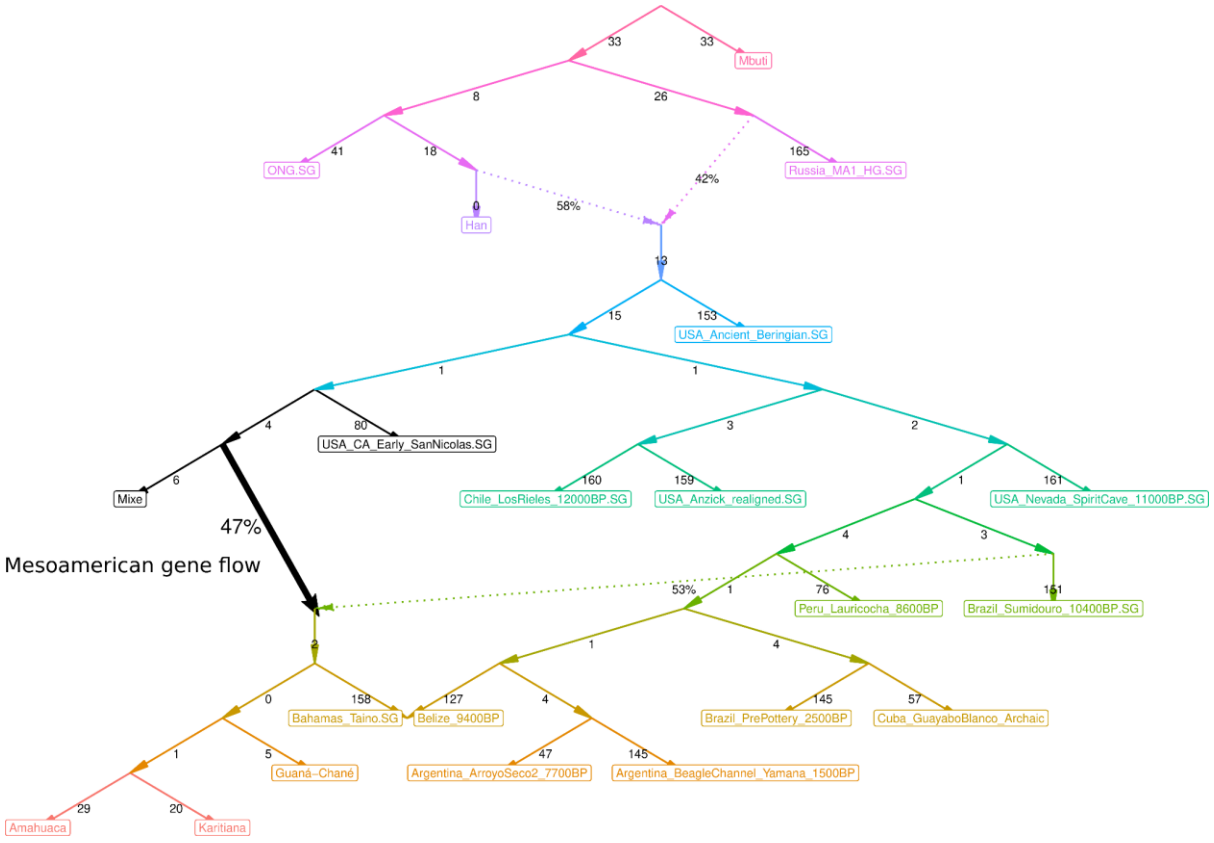
> Present-day Tehuelche and Quechua, which exhibited significant genetic continuity with their second-dispersal ancestors (Figure 4; Extended Data Figure 8C–D), were then included in the admixture graph models to assess their ancestry and genetic continuity. Without admixture events, both clustered together with other contemporary populations and ceramic period Caribbeans (third dispersal), whereas all other Central and South Americans formed a separate branch (second dispersal; Extended Data Figure 10D). Taking admixture events into account, the Tehuelche show genetic continuity with second-dispersal ancestral populations from the Southern Cone (Argentine Pampas), but the majority of their ancestry originates from a sister branch of Chaco populations (third dispersal). In contrast, the Quechua derive most of their ancestry from the earliest-diverging branch of third-dispersal populations, while also exhibiting genetic continuity with second-dispersal ancient Andean individuals (Extended Data Figure 10F–G).

Comment 5. Another key point to add to the "third-wave" discussion, especially concerning the admixture model in line 451, is a discussion of the limitations of using modern "Mixe" as a source. Using a current population introduces a weakness in the model because, due to the lack of ancient DNA data, this issue cannot be fully addressed. However, providing a brief overview of the population history of this source and explaining why the authors believe it is a valuable reference would be helpful.

Authors' reply: We understand the reviewer's concern and acknowledge that the rationale (as well as the limitations) of using the Mixe as the representative Mesoamerican population was not previously explained in the text. The Mixe were selected because they allow us to account for the ancestry contribution of UPopA, which was modeled as an unsampled source

contributing to the ancestry of some present-day Mesoamericans (Moreno-Mayar et al. 2018). This is reflected in our summary admixture graph model presented in Figure 5A, where UPopA is labeled as PopA next to the corresponding edge.

Our summary admixture graph model supports a population history in which UPopA contributed ancestry to both present-day Mesoamericans and ancient individuals from California, providing a good fit to the genetic data. In this context, both our automated admixture graph approach (Extended Data Figure 10C) and the summary model (Figure 5A) indicate that gene flow into third-dispersal Indigenous South American and Caribbean populations from a group closely related to both present-day Mesoamericans and ancient individuals from the Pacific Coast of North America (represented by the Mixe and Early San Nicholas individuals, respectively) is consistent with the observed genetic variation. Importantly, this means that we are not modeling the gene flow into these populations using only the Mixe as proxies for the source. Instead, the model incorporates both the Mixe and ancient individuals from the California Channel Islands as proxies for the ancestral source that contributed to these populations. For clarity, Extended Data Figure 10C is reproduced below highlighting (in black) the Mesoamerican gene flow and the proxies used for the ancestry source.



We have revised the relevant section to more clearly state that including the Mixe population in the admixture graph model was essential to account for the ancestry contribution from the unsampled population A (Figure 5A). We also clarified that both present-day Mixe individuals and ancient individuals from the California Channel Islands share a close genetic relationship

with the ancestry source contributing gene flow to third-dispersal populations. Finally, we acknowledge that incorporating data from ancient individuals from Mesoamerica would enable a more precise determination of the spatiotemporal origin of this ancestry component.

> We integrated our findings with the existing body of knowledge about population history involving the Indigenous American individuals to construct a comprehensive admixture graph model (Methods; Figure 5). In this summary model, we also incorporated the ancestry contribution of the unsampled population A (denoted as PopA in Figure 5A), which was modeled as a source contributing to the ancestry of present-day Mixe individuals[5]. Our results indicate that a population history in which this unsampled population contributed to the ancestors of both present-day Mesoamericans and ancient individuals from California provides a good fit to the genetic data. It is noteworthy that gene flow into the third-dispersal populations was not modeled using only the Mixe as proxies for the source. Rather, the model incorporates both the Mixe and ancient individuals from the California Channel Islands as a sister branch for the ancestral source contributing to third-dispersal populations.

Comment 6. Overall, this is an important and exciting study. The new data presented is invaluable, and the analytical results provided enhance our understanding of the evolution of Indigenous South American genomic diversity. Most of my critique here is minor and based on the approach I would take as an Anthropological Geneticist in describing the data and contextualizing the observations. None of this should prevent the publication of this excellent work. I recommend accepting this manuscript for publication with minor revisions.

Authors' reply: Once again, we thank the reviewer for his insightful comments and constructive criticism. We greatly appreciate the thoughtful feedback, which has helped improve the quality of our manuscript.

References

Bergström, Anders, Shane A. McCarthy, Ruoyun Hui, Mohamed A. Almarri, Qasim Ayub, Petr Danecek, Yuan Chen, et al. 2020. “Insights into Human Genetic Variation and Population History from 929 Diverse Genomes.” *Science* 367 (6484). https://doi.org/10.1126/science.aay5012.

Mallick, Swapan, Heng Li, Mark Lipson, Iain Mathieson, Melissa Gymrek, Fernando Racimo, Mengyao Zhao, et al. 2016. “The Simons Genome Diversity Project: 300 Genomes from 142 Diverse Populations.” *Nature* 538 (7624): 201–6.

Maples, Brian K., Simon Gravel, Eimear E. Kenny, and Carlos D. Bustamante. 2013. “RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference.” *American Journal of Human Genetics* 93 (2): 278–88.

Moreno-Mayar, J. Víctor, Lasse Vinner, Peter de Barros Damgaard, Constanza de la Fuente, Jeffrey Chan, Jeffrey P. Spence, Morten E. Allentoft, et al. 2018. “Early Human Dispersals within the Americas.” *Science* 362 (6419). https://doi.org/10.1126/science.aav2621.

Nunes, Kelly, Marcos Araújo Castro E Silva, Maíra R. Rodrigues, Renan Barbosa Lemes, Patricio Pezo-Valderrama, Lilian Kimura, Lucas Schenatto de Sena, et al. 2025. “Admixture’s Impact on Brazilian Population Evolution and Health.” *Science (New York, N.Y.)* 388 (6748): eadl3564.

Skoglund, Pontus, Swapan Mallick, Maria Cátira Bortolini, Niru Chennagiri, Tábita Hünemeier, Maria Luiza Petzl-Erler, Francisco Mauro Salzano, Nick Patterson, and David Reich. 2015. “Genetic Evidence for Two Founding Populations of the Americas.” *Nature* 525 (7567): 104–8.

Villa-Islas, Viridiana, Alan Izarraras-Gomez, Maximilian Larena, Elizabeth Mejía Perez Campos, Marcela Sandoval-Velasco, Juan Esteban Rodríguez-Rodríguez, Miriam Bravo-Lopez, et al. 2023. “Demographic History and Genetic Structure in Pre-Hispanic Central Mexico.” *Science (New York, N.Y.)* 380 (6645): eadd6142.

RESPONSE TO REVIEWERS

REVIEWER 1

Comment 1. We previously noted that the section "Indigenous American genomes reveal higher-than-expected novel diversity" was misleading. The authors argued that we misinterpreted what the expectation was, but I do not think this is the case. For example, the manuscript says "our results revealed a higher-than-expected genetic diversity compared to that of other continental populations". It later says "By analyzing 128 high-coverage genomes …, we identified 12,493,650 single nucleotide variants (SNVs), of which 11.4% were novel (1,426,511; Figure 1B)." In order to be consistent with the first quoted text, other continental populations should have a lower genetic diversity than this. Instead, the paragraph continues with "African populations exhibit approximately 5M (million) novel variants in 180 individuals17, while Oceanian populations show approximately 2M novel variants in 159 individuals." Clearly there are fewer novel variants among the Indigenous American genomes than Oceanian populations (1.4M vs 2M) with only a small change in sample size. The first quote is therefore inaccurate.

Authors' reply: We appreciate the reviewer's comment and their effort to ensure the accuracy of our conclusions. In response, we have restructured and simplified this section, removing statements about higher-than-expected genetic diversity and limiting it to reporting the observed number of new variants. The new section is:

> **Global genomic databases miss extensive Indigenous American diversity**
>
> Due to an evolutionary history marked by both ancient (out-of-Beringia[3]) and recent (European colonization[11]) demographic shifts, Indigenous populations from America have traditionally been considered and expected to be less genetically diverse than other continental human groups[1,12,13]. By analyzing 128 high-coverage genomes against a dataset of more than 270,000 individuals from various genomic databases (1KGP[21], HGDP[13], SGDP[12], GNOMAD[22], and dbSNP[23]), we identified 12,493,650 single-nucleotide variants (SNVs), of which 11.4% were novel (1,426,511; Figure 1B). Among these new variants, 4.7% were polymorphic (allele frequency (AF) > 0.01), and 0.02% were common (AF > 0.05; Figure 1C-E). For comparison, genomic datasets from other continents with similar sample sizes report greater numbers of novel variants than those observed in Indigenous Americans (~11.1 thousand new variants per individual). Recent studies indicate that African populations carry around 5 million novel variants in 180 individuals (~27.8 thousand new variants per individual)[17], while Oceanian populations exhibit roughly 2 million novel variants in 159 individuals (~12.6 thousand new variants per individual)[18]. Both studies included diverse population sets, and the estimates of novel variants were obtained using the same reference genomic databases as in the present study.

Despite having lower overall genetic diversity than other continental groups, Indigenous Americans harbor a large number of previously unreported variants. This extensive set of novel variants likely reflects both a substantial genetic diversity and the historical underrepresentation of Indigenous Americans in genomic datasets, highlighting the need to include these populations more systematically in genomic research. Greater inclusion will not only deepen our understanding of evolutionarily and biomedically relevant human variation but also help ensure that Indigenous and descendant communities benefit from the outcomes of genomic studies. Although we did not assess structural variation here, these genomes are likely to reveal an additional layer of previously uncharacterized structural diversity, including copy number variation and insertions and deletions.

Comment 2. The model that the authors use to calculate an expected number of "~936,000 expected new variants" is based on an inaccurate model of Indigenous Americans. It does not appear to take into account the known population structure among all the groups that have existed since the Beringian bottleneck. The authors do make several claims about the bottlenecks and founder events that Indigenous Americans have faced, but do not at all comment on the rapid census expansions within the Americas. Regardless, if the authors developed an accurate demographic model of the Indigenous American populations, then they would recapitulate the observed number of genetic variants observed in the populations. Seeing a difference between observed data an the expectation of a model simply means the authors are using a poor model. Perhaps consider developing an improve demographic history using a tool like momi2 (https://www.tandfonline.com/doi/full/10.1080/01621459.2019.1635482).

Authors' reply: We agree with the reviewer that our model does not accurately capture the population history of Indigenous Americans, and our intention was simply to provide a baseline for comparison. However, as noted in our previous response, we have decided to simplify this section and remove the discussion of expected genetic variation. We now only report the number of new variants identified in the newly sequenced genomes.

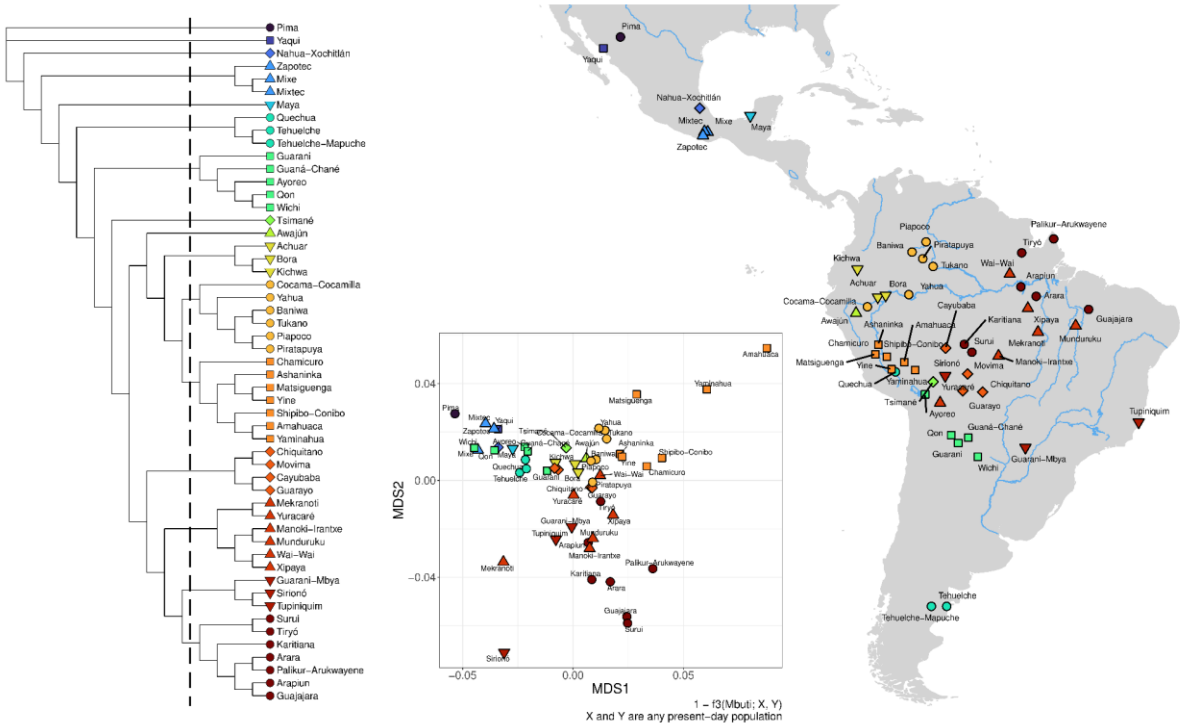
Comment 3. Generally speaking, I do not think that genetic variation should be a bake off, whereby a population is valued because it contains more or less genetic variation than other populations. As proposed, the premise of this entire section seems somewhat troubling to me.

Authors' reply: We agree with the reviewers' comments. As noted earlier, we now avoid discussing expected levels of genetic diversity and instead focus solely on reporting the number of new variants identified in the newly sequenced genomes.

Comment 4. It does not seem like "Indigenous American genetic variation mirrors geographic distribution" is an appropriate section title for that section. While the PCA plot in Figure 2 does show the Pima at the top (representing "North"), it is not at all clear that that the vast majority of points follow the geographic pattern. For example, there are many groups sampled to the south of Surui, east of the Karitiana, and west and south of Amahuaca and Yaminahua. Yet these populations end up as the major labeled outliers. The neighbor joining tree in Figure 3 does not appear to convey geography. To argue that there is a geographic similarity in any of these plots, perhaps follow what was done in the Novembre et al (2008) paper that did this kind of analysis in Europeans? To this reviewer, it looks like the peopling of the Americas was a much more complicated process than the radiation in Europe, and potentially much more exciting to describe the complexity than to try to narrow it down to simple geography. Statements like "Analysis of molecular variance (AMOVA) suggested that genetic clusters corresponding to broad geographic regions accounted for a significant proportion of genetic variation (p = 0.04)" are not convincing. One of the newly added sentences captures my sentiment: "However, an isolation-by-distance model alone may not fully explain the observed genetic variation across the region, as additional factors are likely contributing to the observed variation."

Authors' reply: We agree with the reviewer that geographic distribution alone cannot fully explain the observed patterns of genetic variation, as already acknowledged in the manuscript. As discussed therein, the structure observed in the PCA (Figure 2A) is partly, and possibly largely, driven by the elevated levels of inbreeding experienced by some populations, as reflected by their high ROH-based inbreeding coefficients (Figure 2B). However, it is important to note that both PCA and ADMIXTURE inferences are strongly influenced by population-specific genetic drift, which can obscure the underlying genetic structure among populations. Nonetheless, the PCA and ADMIXTURE results still reveal at least a modest level of genetic differentiation among populations from distinct geographic regions.

Additional analyses, particularly outgroup f3 statistics that provide a measure of genetic similarity that is unbiased by population-specific drift, indicate that geography accounts for a substantial part of the observed genetic variation, especially at broad spatial scales. Clear differentiation is observed between North and South America and among regions, including the Andes, Southern Cone, Chaco, Amazon, Aridoamerica, and Mesoamerica. These patterns are consistently recovered across independent analyses of present-day populations, including neighbor-joining trees and multidimensional scaling plots based on outgroup f3 statistics (Figure 3A,C). Even at finer levels of clustering, genetically similar populations are generally geographically proximate (see figure below). This correspondence remains when ancient individuals are included, as shown by ADMIXTURE (Extended Data Figure 9) and outgroup f3 statistics (Figure 4A,B), reflecting a degree of genetic continuity.



Our use of AMOVA was intended to formally test the association between geographic distribution and genetic variation, and the results show a significant association ($p < 0.05$). Additional support comes from the Spearman correlation coefficients we estimated for genetic (measured as pairwise 1–outgroup f3) and geographic distances, both between pairs of present-day populations and between present-day and ancient populations, which revealed significant positive correlations in both cases ($p < 0.01$; Supplementary Figures 3 and 4).

In addition to these findings, the estimated effective migration surface analyses indicate that several regions show genetic differentiation greater than expected under an isolation-by-distance model, particularly southeastern Peru, northern Bolivia, and southwestern Amazonia. This pattern further supports the conclusion that factors beyond geographic distance have contributed to shaping the observed distribution of genetic variation.

We appreciate the reviewer's concern that the original section title did not clearly communicate these nuances, so we have revised it to better reflect that, although geography explains a statistically significant portion of the genetic variation, it does not account for all of it. The new title is:

Demographic history and geography leave distinct signatures in genetic variation

We also revised this section to clarify the limitations of the PCA and ADMIXTURE analyses and to explain why genetic differentiation among populations from distinct geographic regions may be less evident in these results. In addition, we provide a more detailed explanation of the proportion of genetic variation explained by the AMOVA.

We characterized present-day genetic structure and variation among Indigenous American populations using Principal Component Analysis (PCA) and ADMIXTURE (see Methods) on a masked dataset of 160 unrelated genomes. PCA of global reference populations revealed no evidence of batch effects, as individuals were consistently clustered by geographic origin rather than by dataset source (Supplementary Note 1). Our analyses revealed genetic differentiation between North and South American Indigenous groups, as well as substructure within each region (Figures 2 and 3). PCA and ADMIXTURE inferences are strongly influenced by population-specific genetic drift[24]. Accordingly, several populations, including the Karitiana, Suruí, Amahuaca, and Yaminahua, display highly distinctive genetic profiles. These patterns are driven at least in part by elevated levels of inbreeding, as indicated by the larger proportion of the genome contained within runs of homozygosity (ROH), measured by F_{ROH}, an ROH-based inbreeding coefficient (Figure 2; Extended Data Figure 3; Supplementary Table 4). These results underscore the limitations of using the Suruí and Karitiana as representative proxies for lowland South American genetic diversity (a common practice in population genomic studies), given their pronounced genetic differentiation, even relative to geographically nearby populations.

The relationship between genetic similarity and geographic distribution becomes clearer when examined using outgroup f3 statistics, which provide a measure of genetic similarity that is unbiased by population-specific genetic drift. A neighbor-joining (NJ) tree and multidimensional scaling (MDS) applied to 1–outgroup f_3 and 1/outgroup f_3 (Methods; Figure 3; Extended Data Figure 8A–B; Supplementary Table 5) revealed four major genetic clusters within South America that align closely with geographic regions. Accordingly, we named them based on their respective locations: (i) Southern Cone, (ii) Eastern, (iii) Western (mostly lowlands) South America, and (iv) Chaco. The Southern Cone populations (Mapuche-Tehuelche) were the most differentiated in South America, yet the Andean Quechua cluster was positioned within the same clade, suggesting a long-distance genetic connection. The Chaco populations form the sister branch, followed by Tsimané, which remains unclustered, possibly because of their cultural and linguistic isolation[24,25]. The final division separates Eastern and Western South American lowland populations.

[...]

Analysis of molecular variance (AMOVA) grouping individuals by genetic clusters explained a fraction of the total genetic variation (~9%; $p = 0.04$). In contrast, grouping individuals by ethnolinguistic units (language families) did not explain a detectable proportion of the variation (~0%; $p = 0.57$; Supplementary Table 6). Additionally, genetic distances, measured as $1–f_3$(Mbuti; X, Y), are significantly correlated with great-circle geographic distances between all pairs of present-day Indigenous American populations (Spearman's $r = 0.4401$, $p \cong 0$ for all Indigenous Americans; $r = 0.148$, $p \cong 0$ for South America alone; Supplementary Table

7 and Supplementary Note 4). However, an isolation-by-distance model alone may not fully explain the observed genetic variation across the region, as additional factors are likely contributing to the observed variation. In line with this, the estimated effective migration surface (EEMS) models indicated regions with genetic differentiation greater than expected under an isolation-by-distance model, notably southeastern Peru, northern Bolivia, and southwestern Amazonia (Extended Data Figure 4). This pattern closely resembles findings from a previous study involving a more limited set of populations from western South America[29]. Lower inferred migration rates were observed in Aridoamerica, the northern Peruvian and Ecuadorian Amazon, the Chaco region, and eastern South America, especially in areas north of the Amazon River. These patterns suggest that demographic history and geographic and cultural factors (including topography, climate, river systems, vegetation density, and sociocultural boundaries) beyond geographic distance have also shaped genetic differentiation. Additionally, an unexpected signal of increased migration rate between Mesoamerica and South America indicates greater genetic similarity than geographic distances alone would predict. This observation aligns with previous findings indicating gene flow from North and Central American populations into South American groups[5,33].

Comment 5. The authors seem to neglect the importance of the Gusareva paper. While in the rebuttal the authors claimed a difference in the number of populations sampled, they did not notice that several of the populations included in the present study were pooled into single groups in the Gusareva study. I think it would be helpful to acknowledge how the current study amplifies and clarifies some of the messages of the Gusareva et al paper and adds new insights, rather than attempt to belittle the paper with terms like "only".

Authors' reply: Thank you for your comment. We apologize if our response letter gave the impression that we were downplaying the work of Gusareva et al. This was never our intention. Our goal was solely to emphasize that the two studies differ substantially in their scope and focus, particularly with respect to sampling resolution and the level of anthropological knowledge available for the populations sampled and analyzed. In line with this, Gusareva et al. are cited throughout our manuscript where relevant, and the points of convergence and divergence between the studies are clearly indicated and discussed.

REVIEWER 2

I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

Authors' reply: We thank the reviewers for their comments, which have significantly improved our manuscript.

REVIEWER 3

I am satisfied with the response to most of my comments. I have very minor things that I think still require some attention.

Comments:

Comment 1. Both reviewer 1 and I raised concerns about the model to test whether the novel variation identified is a lot or "expected". In my opinion, the generation of this genomic data is an extremely valuable contribution on its own right, and an analysis to justify it is not needed (at least not in the main text). I think that stating that a lot of new variation is found is good enough. The space used to contrast against a given model (that could be in principle parametrized in many ways) is needless and distracts from the many other analyses and inferences that come later in the text. This is just a suggestion to be gentle on the readers and not saturate them beyond the minimum necessary.

In any case, if the authors decide to keep this part, along with the new additional information produced in response to Reviewer 1, then more discussion should be given as to why this is relevant. Is it because the demographic model is wrong? Or why?

Authors' reply: We appreciate the reviewer's helpful and insightful comments, which we believe have significantly improved our manuscript and ensured the accuracy of the conclusions drawn from our results. In response to the reviewers' suggestions, we have restructured and simplified this section. We now focus exclusively on reporting the number of new variants identified in the newly sequenced genomes, rather than discussing expected levels of genetic diversity.

Comment 2. Reviewer 4 and I agree that the Meso -> South geneflow is not unexpected. I appreciate the added text to contextualize, but I think removing the word "unexpected" is still needed.

Authors' reply: We agree with the reviewer's point and have removed the word "unexpected" from this sentence.

Comment 3. The addition of a closing sentence to the section "Uneven genetic diversity distribution and the lasting impact of colonization" helps in giving more sense to the title. However, the text throughout the section does not seem to assert a cause-consequence (i.e. "impact") relationship. I think I missed that the authors were trying to, in fact, state the "lasting impact of colonization" because this was not explicitly or actively described as such.

Authors' reply: We thank the reviewer for raising this issue. We agree that it is essential not only to demonstrate the impacts of European colonization on Indigenous American and Latin American populations, but also to clearly explain to the reader the cause-and-effect relationship between these impacts and colonization. We acknowledge that this connection was not always explicitly or directly presented in several parts of the text, as the reviewer also notes in the following comments. In addition, we did not clearly articulate the often catastrophic and violent impacts experienced by Indigenous American populations as a result of European colonization. To address this point, we have now explicitly linked colonization processes to the patterns of present-day genetic diversity throughout the section "Uneven genetic diversity distribution and the lasting impact of colonization." Specific modifications are detailed in our responses to the comments below.

Drawing from the sentences highlighted in the response by the authors:

Comment 4. "These groups also exhibited an excess of long ROH segments (Extended Data Figure 5), indicating recent inbreeding likely caused by population collapse, fragmentation, and post-colonization isolation" -> What caused this collapse?

Authors' reply: As noted in our response to comment 3, we have now explicitly linked the colonization processes to the patterns of present-day genetic diversity described throughout the manuscript. The revised paragraph is as follows:

> Compared to other continental populations, Indigenous Americans exhibit a distinctive length distribution of runs of homozygosity (ROH), with a higher average number of segments across all length categories, except for the shortest ones (< 2 cM), where Oceanians show a greater number (Extended Data Figure 5), indicating both strong ancient population bottlenecks and recent inbreeding. The Mesoamerican and Southern Cone populations showed the lowest ROH levels among Indigenous Americans (Extended Data Figs. 3 and 5), suggesting greater genetic diversity. In contrast, Moseten (Tsimané), Pano-Takana (Amahuaca and Yaminahua), Zamuco (Ayoreo), and Tupi (e.g., Sirionó, Suruí, and Karitiana) have the highest ROH counts and total length (Extended Data Figs. 3, 5; Supplementary Table 4). These groups also showed an excess of long ROH segments (Extended Data Figure 5), likely resulting from population collapse, fragmentation, and isolation following European colonization, as previously shown[11,33,34]. Contributing factors included epidemics, enslavement, wars of conquest, forced displacement, the instigation of intergroup violence, habitat destruction, and the disruption of subsistence strategies and traditional knowledge[35–38].

Comment 5. ”Additionally, identity by descent (IBD) analyses revealed that segments >8 cM were predominantly shared within populations (Extended Data Figure 6; Supplementary Table 8), suggesting limited gene flow since the colonial period. Therefore, these small and isolated populations likely experienced increased inbreeding during this period, leading to the high prevalence of long ROHs (Extended Data Figure 5).” -> What caused this limited gene flow?

Authors' reply: The new paragraph is:

> Additionally, identity-by-descent (IBD) analyses revealed that segments >8 cM were predominantly shared within populations (Extended Data Figure 6; Supplementary Table 8), suggesting limited gene flow since the colonial period. Therefore, these small and isolated populations likely experienced increased inbreeding during this period, leading to the high prevalence of long ROHs (Extended Data Figure 5). This pattern again reflects the extermination and displacement of Indigenous American communities caused by European colonization, which fragmented populations, reduced intergroup contact, and increased mating within groups.

Comment 6. ”Effective population size (Ne) histories inferred from IBD analysis (methods), grouping individuals by genetic cluster, revealed a widespread post-European contact bottleneck across all Indigenous American populations, followed by signs of recent genetic diversity recovery in Chaco and western South America (Figure 3D).” -> What caused this “bottleneck”?

Authors' reply: This section of the text has been revised as follows:

> Further evidence of the impact of European colonization comes from effective population size (Ne) histories inferred from IBD analysis (Methods). When individuals were grouped by genetic cluster, these histories revealed a widespread post-European-contact bottleneck across all Indigenous American populations, followed by indications of a recent recovery of genetic diversity in Chaco and western South America (Figure 3D).

Comment 7. “In contrast, Aridoamerica exhibits persistently low Ne values and experienced the most pronounced post-European contact bottleneck (Figure 3D), consistent with the high amount of long IBD segments shared within this cluster (Extended Data Figure 6).” -> same as above.

Authors' reply: This section was also revised as follows:

> In contrast, Aridoamerica shows persistently low effective population sizes and underwent the most severe bottleneck following European contact (Figure 3D),

reflecting a stronger impact of colonization in the region. This pattern is consistent with the high number of long IBD segments shared within this cluster (Extended Data Figure 6).

Comment 8. I am certain that authors are aware and sensible to the lasting impact of colonial dominance on Indigenous Americas, the passages above, however, don't convey it. Referring to the Indigenous population collapse as a passive "demographic bottleneck" sanitizes history and obscures the known causes: epidemics, massacres, and displacement directly perpetrated by colonizers. The text as it stands reports findings accurately and soundly, but my very personal opinion is that historical accountability is important. This is just a personal recommendation and authors can decide what is best.

Authors' reply: We thank the reviewer for this comment and fully agree. Our intention was not to sanitize the history of Indigenous Americans, but we acknowledge that the presentation of the results did not clearly link the observed patterns of genetic diversity to their historical context.

We totally agree that it is essential to communicate this clearly to readers. Highlighting these points underscores the historical processes of colonization that have shaped and continue to influence the genetic diversity of both Indigenous and non-Indigenous American populations. It is equally important to stress that the power imbalances established during colonization persist today, including the underrepresentation of these populations both as participants in genomic studies and as researchers leading such work. These disparities will limit the ability of these populations to fully benefit from the medical advances, economic opportunities, and social development offered by genomic research. This context underscores the importance of the present work, which addresses the gap in representation of Indigenous Americans and was conducted by a team composed almost entirely of Latin American researchers.

REVIEWER 4

The authors present a revised version of their manuscript, featuring 128 new, high-coverage Indigenous American genomes from 45 populations and 28 language families, distributed across several Central and South American countries. These new data are an invaluable resource for a continent and its peoples that have been seriously understudied, and the presented results represent a significant advancement in our understanding of South American population history. In my previous review of the study, I voiced some minor concerns regarding the contextualization of the authors' analyses with previously published studies and some minor issues with clarity. The revised manuscript effectively addressed all these critiques, and, along with the revisions based on other reviewers' comments, this manuscript has undergone substantial improvement. The authors now present their excellent analytical work in a comprehensive and thoughtful manner. I am very excited about this important work and recommend the manuscript for publication without further revisions.

Authors' reply: We are very grateful for the reviewer's comments, which we believe have substantially strengthened the clarity and quality of our manuscript.

RESPONSE TO REVIEWERS

REVIEWER 1

Overall:

I am honored to have been a part of this review process. I think the revisions suggested by all reviewers have substantially strengthened manuscript, with all claims in the paper now being strongly supported with direct analyses, or toned down to reflect observations in the data. The work was conducted in collaboration with Indigenous communities, and data were shared back to the communities. This is an underappreciated, yet essential step for representative genomics projects that all studies should emulate. I have a few remaining minor comments that in my view should not hold up the manuscript with further review.

Section "Global genomic databases miss extensive Indigenous American diversity"

Comment 1. I would recommend removing the sentence "Although we did not assess structural variation here, these genomes are likely to reveal an additional layer of previously uncharacterized structural diversity, including copy number variation and insertions and deletions." from the section. It is too speculative for a results section, and could be moved to the conclusions if the authors want to keep it.

Authors' reply: We appreciate the reviewer's thoughtful considerations and comments throughout this review process. Following the reviewer's suggestion, we have moved the referred sentence to the "Conclusion" section of the manuscript:

> Our comprehensive genomic analysis of 128 high-coverage Indigenous American genomes across 45 ethnic groups from eight Latin American countries revealed a complex and dynamic demographic and evolutionary history that improved our understanding of human diversity and adaptation in the American continent. We present evidence of at least three major population dispersals into South America, with subsequent regional differentiation and long-term genetic continuity in many areas, challenging oversimplified models of continental peopling. We demonstrated that Indigenous American genomes harbor extensive novel genetic diversity, highlighting the need to better represent these populations using global genomic datasets. This diversity reflects both deep-time evolutionary processes, such as the Beringian standstill and post-glacial expansions, and more recent demographic shifts, including severe bottlenecks following European colonization and subsequent population isolation. Although structural variation was not assessed, these genomes likely contain substantial uncharacterized diversity, including copy number variation

and insertions and deletions. Our analyses revealed widespread signatures of natural selection across loci related to immune function, metabolism, fertility, and development, suggesting that Indigenous Americans underwent strong and diverse selective pressures shaped by the continent's vast environmental heterogeneity. Particularly remarkable are genomic regions with excess allele sharing with Australasian populations, the Ypykuéra signal, many of which we show are likely targets of positive selection. This suggests that the persistence of this genetic signal, present at low but consistent levels for over 10,000 years, is likely not just a signal of ancient gene flow but may also reflect adaptive advantages in specific environmental contexts. Furthermore, we identified candidate regions of adaptive archaic introgression from Neanderthals and Denisovans that contribute to functions related to immunity, metabolism, and epidermal integrity, thereby reinforcing the role of archaic alleles in shaping the evolutionary trajectory of non-African populations. Importantly, our data indicate minimal overlap between genomic regions with Australasian affinity and those introgressed from archaic hominins, supporting the interpretation that these signals represent distinct evolutionary phenomena. Together, these findings underscore the unique evolutionary history of Indigenous Americans, shaped by complex interactions between ancient population structure, migration, natural selection, and introgression. They also emphasized the importance of expanding genomic research among underrepresented populations to capture the full scope of human genomic diversity.

Comment 2. To me, the new text of this section jumps back and forth between two topics. I would recommend moving the sentence "By analyzing 128 high-coverage genomes against a dataset of more than 270,000 individuals from various genomic databases (1KGP21, HGDP13, SGDP12, GNOMAD22, and dbSNP23), we identified 12,493,650 single nucleotide variants (SNVs), of which 11.4% were novel (1,426,511; Figure 1B). Among these new variants, 4.7% were polymorphic (allele frequency (AF) > 0.01), and 0.02% were common (AF > 0.05; Figure 1C-E)." down to the second sentence of the second paragraph where the authors focus on novel variation, so that the first paragraph just talks about total variation and the second paragraph just talks about novel variation.

Authors' reply: We appreciate the reviewer's comments regarding the organization of this section. After careful consideration, we believe that the current structure offers the clearest and most logical flow for the reader. Presenting the results first, in the first paragraph, allows the observations to be clearly established before discussing their implications in the second paragraph. For this reason, we have decided to maintain the current paragraph structure.

Section "Demographic history and geography leave distinct signatures in genetic variation"

Comment 3. The conclusion that Suruí and Karitiana may be poor representatives for South American Indigenous ancestry in the new text is important and clearly laid out.

Authors' reply: We again thank the reviewer's constructive suggestions.

REVIEWER 2

Comment 1. I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

Authors' reply: We once again thank all the reviewers for their careful attention and helpful suggestions throughout the review process.

REVIEWER 3

Comment 1. I have no further comments and appreciate the efffort by authors to address the feedback.

Authors' reply: We once again thank all the reviewers for their careful attention and helpful suggestions throughout the review process.